Почтовая система RuPost

Предпочтительная архитектура развертывания (Preferred Architecture)

Версия 1.2 - Май 2026

Официальный веб-сайт: http://www.rupost.ru .

Оглавление

1. Что такое “Предпочтительная архитектура”

Предпочтительная архитектура развертывания или **ПА** (Preferred Architecture, PA) — это совокупность рекомендаций инженеров и разработчиков команды РУПОСТ по построению архитектуры почтовой системы, размещенной как в одном, так и в нескольких территориально разнесенных ЦОД.

Предпочтительная архитектура развертывания почтовой системы RuPost построена с учетом накопленного **опыта развертывания RuPost**, обширной практики применения **предпочтительной архитектуры (preferred architecture) Microsoft Exchange и организации сосуществования и миграции с Exchange на RuPost**. При том, что внутренняя архитектура и технологии RuPost обладают принципиальными отличиями от Exchange, ряд аспектов архитектуры развертывания и требований к инфраструктуре этих систем схожи и являются следствием требований к обеспечению высокой доступности почтовой системы для корпоративных пользователей. Для клиентов, использующих в данный момент серверы Microsoft Exchange, развернутые в соответствии с Exchange PA в нескольких ЦОД, применение представленных рекомендаций позволит существенно облегчить сосуществование и миграцию пользователей.

Данная архитектура разработана с учетом ключевых бизнес-требований корпоративного сектора:

- Высокий уровень доступности системы при развертывании в одном центре обработки данных (ЦОД) или в нескольких ЦОД
- Поддержка реплик почтовых хранилищ и баз данных для резервирования данных и быстрого реагирования на возможные сбои
- Повышение доступности за счет резервирования узлов обработки данных и снижения сложности системы в целом
- Снижение совокупных затрат на инфраструктуру обмена сообщениями

Предпочтительность архитектуры означает, что не каждый клиент сможет развернуть ее “один в один”. Например, не у всех клиентов есть несколько центров обработки данных. Некоторые из наших клиентов могут иметь разные бизнес-требования или внутренние политики соответствия, что может потребовать изменений в архитектуре развертывания. Даже в этих случаях, по-прежнему, можно получить преимущества, если придерживаться предпочтительной архитектуры и отклоняться от нее только в крайних случаях, когда внутренние требования и политики все же заставляют осуществить конкретное развертывание с отличиями от рекомендуемого.

Данное описание предпочтительной архитектуры развертывания RuPost включает описания и комментарии по работе Exchange в справочных целях и для облегчения понимания применяемых в RuPost лучших практиках и технологических решениях, как для тех специалистов, кто ранее работал с Microsoft Exchange, так и для широкого круга инженеров, отвечающих за построение современной масштабируемой отказоустойчивой системы электронной почты корпоративного класса.

Данный документ описывает текущие и готовящиеся к выпуску функциональные возможности решения и может изменятся в соответствии с проводимыми изменениями в продукте.

Источники

Exchange 2019 preferred architecture
https://learn.microsoft.com/en-us/exchange/plan-and-deploy/deployment-ref/preferred-architecture-2019

Exchange Team Blog - Ross Smith IV, Principal Program Manager, Exchange Server: The Exchange 2016 PA
https://techcommunity.microsoft.com/blog/exchange/the-exchange-2016-preferred-architecture/604024

2. Проблематика высокой доступности (High-Availability)

Корпоративная почтовая система является критичным сервисом для функционирования организации, так как обеспечивает основные внутренние коммуникации сотрудников и используется для взаимодействия с внешним миром. Такие системы строятся с учетом требований высокой доступности (high-availability, HA).

Требования высокой доступности формулируются в отношении двух аспектов работы системы:

- Непрерывность обработки информации
- Доступность информации

Оба аспекта организационных практик обычно отражают в рамках процессов **Обеспечения непрерывности бизнеса (Business Continuity Management) и восстановления после сбоев (Disaster Recovery)**. Содержание требований и рекомендаций по Business Continuity/Disaster Recovery (BCDR) находится за рамками данного документа и отражены в таких стандартах как ITIL/ITSM/ГОСТ Р ИСО МЭК 20000 (ISO IEC 20000) и COBIT, на основе которых строятся соответствующие регламенты и процедуры конкретной организации. В свою очередь, в данном документе рассматриваются архитектурно-технологические характеристики и возможности организации высокой доступности почтовой системы RuPost, в том числе, с точки зрения сравнения с привычным для многих организаций Microsoft Exchange для сосуществования и миграции с него на RuPost.

Системы высокой доступности, обычно, организуются в виде многоузловых кластеров и групп кластеров с сервисами обработки и хранения информации и обеспечивают следующие возможности:

- **Резервирование хранения** информации с использованием репликации данных и с возможностью их резервного копирования и восстановления.
- **Отказоустойчивость и резервирование обработки** информации за счет возможности оперативного переключения между резервными узлами/кластерами обработки информации и репликами данных без прерывания или с минимальным прерыванием доступа пользователей.
- **Балансировка нагрузки** между узлами/кластерами для повышения скорости работы пользователей с системой и эффективного использования инфраструктурных ресурсов.
- **Распределенность**, в том числе географическая распределенность, как возможность использования системы в различных топологиях и сценариях применения нескольких центров обработки данных.

Данная предпочтительная архитектура RuPost разработана с учетом поддержки широкого спектра требований к обеспечению высокой доступности почтовой системы.

3. Компоненты и отказоустойчивая архитектура почтовых серверов Microsoft Exchange и RuPost

3.1. Архитектура сервера Microsoft Exchange

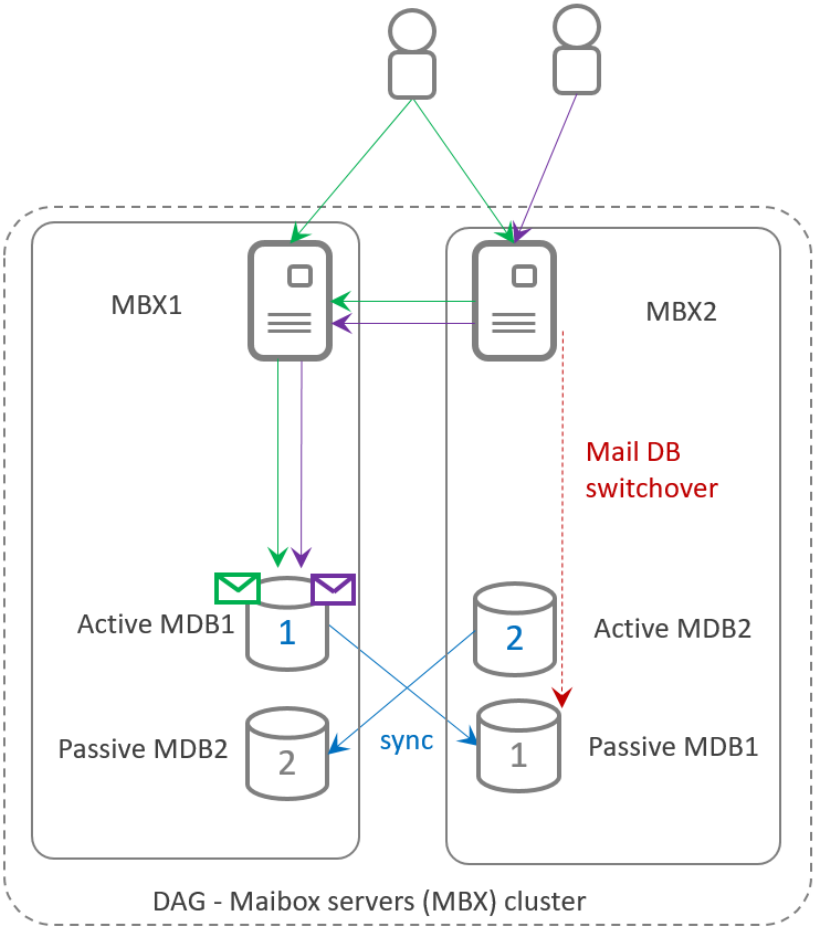
Эволюция сервера Microsoft Exchange, происходившая на протяжении нескольких десятков лет, показывает преимущества и недостатки разных архитектурных подходов, изменявшихся по мере развития версий этой почтовой системы. Например, после периода выделения различных сервисов (в терминах Microsoft – ролей) на разных узлах системы, актуальные версии Exchange объединяют все ключевые функции в рамках одного сервера, называемого Mailbox Server (MBX). В частности, с версий Exchange 2016 серверы MBX стали совмещать и функции обработки клиентских подключений (CAS – Client Access Services) и транспортную службу (Mailbox Transport Service), в том числе проксируя клиентские запросы на другие узлы и группы узлов Exchange. Это было сделано для снижения числа возможных точек отказа и упрощения управления системой. Microsoft позиционирует такую архитектуру как “Integrated design inside each server”.

Для обеспечения высокой доступности, т.е. организации кластера Microsoft Exchange, используется концепция Database Availability Group (DAG). Почтовые ящики пользователей распределяются по почтовым базам (Mailbox Database). Как уже было сказано, отдельные роли, серверы и кластеры CAS теперь не требуются, так как CAS уже является составной частью любого узла кластера актуальных версий Exchange.

В свою очередь, *почтовая система RuPost с самого начала (или как говорят “by design”), проектировалась и разрабатывались в концепции “равнозначных узлов” без выделения специализированных ролей или разных типов узлов*. Это позволило сразу избежать необходимости отдельной кластеризации для разных по своим ролям узлов системы (например, Client Access Array для отказоустойчивости узлов CAS и Database Availability Group для MBX), что качественно снижает сложность архитектуры развертывания почтовой системы и повышает стабильность ее функционирования.

Особенностью архитектуры Exchange является то, что каждый Mailbox Server (MBX) в кластере эксклюзивно работает со своей почтовой базой Mailbox Database (MDB), при условии, что почтовый ящик каждого пользователя может в каждый момент времени обслуживаться только одним MBX. Это означает, что, если в кластерной конфигурации Exchange пользователь через средства балансировки нагрузки подключается не к тому серверу MBX1, который отвечает за работу с почтовой базой, где находится ящик пользователя, а к другому - MBX2, то MBX2 выполняет роль прокси и всегда перенаправляет через себя запросы пользователя на сервер MBX1, который, в свою очередь, уже работает непосредственно с необходимой почтовой базой.

Такой подход требует обеспечения сохранности данных и организации отказоустойчивости для самих почтовых баз данных. Это решается созданием копий (реплик) почтовых баз. При этом, в силу эксклюзивности работы почтового сервера с базой данных, в рамках кластера DAG существует только одна активная почтовая база и ее пассивные копии. Поэтому DAG фокусируется на вопросах создания копий почтовых БД с возможностью ручного и автоматического переключения между экземплярами БД, когда выбирается одна из доступных пассивных копий БД, которая делается активной, и с которой, в дальнейшем, ведется работа как с основной базой данных.

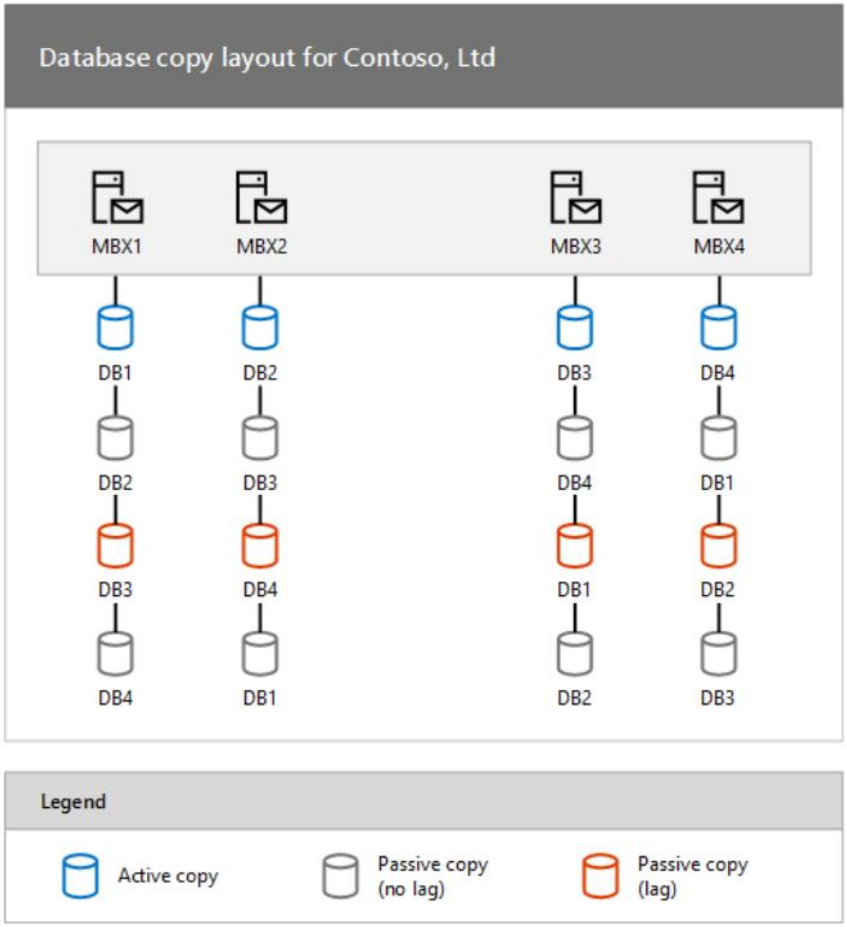


Так как процесс создания целостных и актуальных реплик таких баз данных достаточно ресурсоемкий и требовательный к сетевой инфраструктуре – наравне с пассивными копиями почтовых баз данных Exchange поддерживает и более простые в обслуживании “отстающие” (lagged) копии почтовых баз. В реальных конфигурациях, для снижения стоимости владения, используются как пассивные, так и отстающие копии.

Для организации работы с почтовыми базами и разными системами хранения (JBOD, RAID) Exchange обладает встроенными механизмами работы с почтовыми базами:

- управляемым хранилищем Managed Store, обеспечивающим RPC доступ к ящикам пользователей в почтовой базе;
- сервисом Microsoft Exchange Replication для репликации баз;
- менеджером активации Exchange Server Active Manager, отвечающем за переключение между активной и пассивной репликой базы в случае возникновения сбоев.

Можно сказать, что Exchange, в определенной степени, реализует системные функции аналогичные программно-определяемым хранилищам (SDS, Software Defined Storage) и протоколу NFS для хранения информации, ее репликации и обеспечения доступа через точки монтирования посредством RPC-доступа к заданной реплике хранилища с переключением доступа между репликами в случае возникновения тех или иных сбоев.



Эволюция архитектуры Exchange, даже на современном уровне, оставила ряд вопросов, требующих внимания при организации почтовой системы высокой доступности:

- Эксклюзивность работы экземпляра почтового сервера MBX с почтовой базой данных — с Mailbox database может работать только один почтовый сервер — приводит к несбалансированной нагрузке на разные почтовые серверы, изначально требующей корректного планирования при распределении почтовых ящиков пользователей по разным почтовым базам / серверам.
- Почтовая база, являясь единым хранилищем для почтовых ящиков множества пользователей, требует множества реплик для минимизации рисков, связанных со сбоями чтения-записи на уровне БД или сбоями инфраструктуры, приводящими к недоступности экземпляра почтового сервера MBX:
    - Сбой чтения-записи одного сообщения делает недоступными все почтовые ящики, находящиеся в той же базе данных.
    - Сбой работы одного экземпляра почтового сервера делает недоступными все почтовые ящики пользователей во всех активных почтовых базах данного сервера.
- Число почтовых серверов в кластере DAG ограничено 16 экземплярами, что в больших инсталляциях приводит к необходимости организации множества кластеров DAG, усложняющих управление почтовой системой.
- При организации кластера из N (N<=16) серверов, в силу эксклюзивности работы экземпляра почтового сервера с БД, предполагается создание, как минимум, N почтовых баз данных, каждая из которых должна обладать несколькими репликами для снижения рисков сбоев. Это

приводит к повышению нагрузки на инфраструктурные ресурсы и ресурсы администрирования системы.

RuPost учитывает данные особенности и ограничения Exchange и реализует альтернативный подход в организации почтовой системы высокой доступности, основанный на уникальных архитектурных решениях и системных особенностях используемой ОС Linux.

Источники

History of Microsoft Exchange Server
https://en.wikipedia.org/wiki/History_of_Microsoft_Exchange_Server

Changes to high availability and site resilience over previous versions of Exchange Server
https://learn.microsoft.com/en-us/exchange/high-availability/ha-changes?view=exchserver-2019

Exchange architecture
https://learn.microsoft.com/en-us/exchange/architecture/architecture?view=exchserver-2019

The Exchange 2016 Preferred Architecture
https://techcommunity.microsoft.com/t5/exchange-team-blog/the-exchange-2016-preferred-architecture/ba-p/604024

Client Access services
https://learn.microsoft.com/en-us/exchange/architecture/client-access/client-access?view=exchserver-2019

High availability and site resilience in Exchange Server
https://learn.microsoft.com/en-us/exchange/high-availability/high-availability?view=exchserver-2019

Deploying high availability and site resilience in Exchange Server
https://learn.microsoft.com/en-us/exchange/high-availability/deploy-ha?view=exchserver-2019

Load balancing in Exchange Server
https://learn.microsoft.com/en-us/exchange/architecture/client-access/load-balancing?view=exchserver-2019

Exchange Switchovers and failovers
https://learn.microsoft.com/en-us/exchange/high-availability/manage-ha/switchovers-and-failovers?view=exchserver-2019

Manage mailbox databases in Exchange Server
https://learn.microsoft.com/en-us/exchange/architecture/mailbox-servers/manage-databases?view=exchserver-2019

Manage mailbox database copies
https://learn.microsoft.com/en-us/exchange/high-availability/manage-ha/manage-database-copies?view=exchserver-2019

Managed Store
https://learn.microsoft.com/en-us/exchange/managed-store-exchange-2013-help?view=exchserver-2019

Exchange Server Active Manager
https://learn.microsoft.com/en-us/exchange/high-availability/database-availability-groups/active-manager?view=exchserver-2019

Mail flow and the transport pipeline
https://learn.microsoft.com/en-us/exchange/mail-flow/mail-flow?view=exchserver-2019

Exchange Server 2019 Sizing Calculator
https://aka.ms/Exchange2019Calc
https://www.microsoft.com/en-us/download/details.aspx?id=102123

https://practical365.com/exchange-server-role-requirements-calculator/

Backup, restore, and disaster recovery in Exchange Server - Exchange Native Data Protection
https://learn.microsoft.com/en-us/exchange/high-availability/disaster-recovery/disaster-recovery?view=exchserver-2019

Exchange Server: Use Windows Server Backup to back up and restore Exchange data
https://learn.microsoft.com/en-us/exchange/high-availability/disaster-recovery/windows-server-backup?view=exchserver-2019

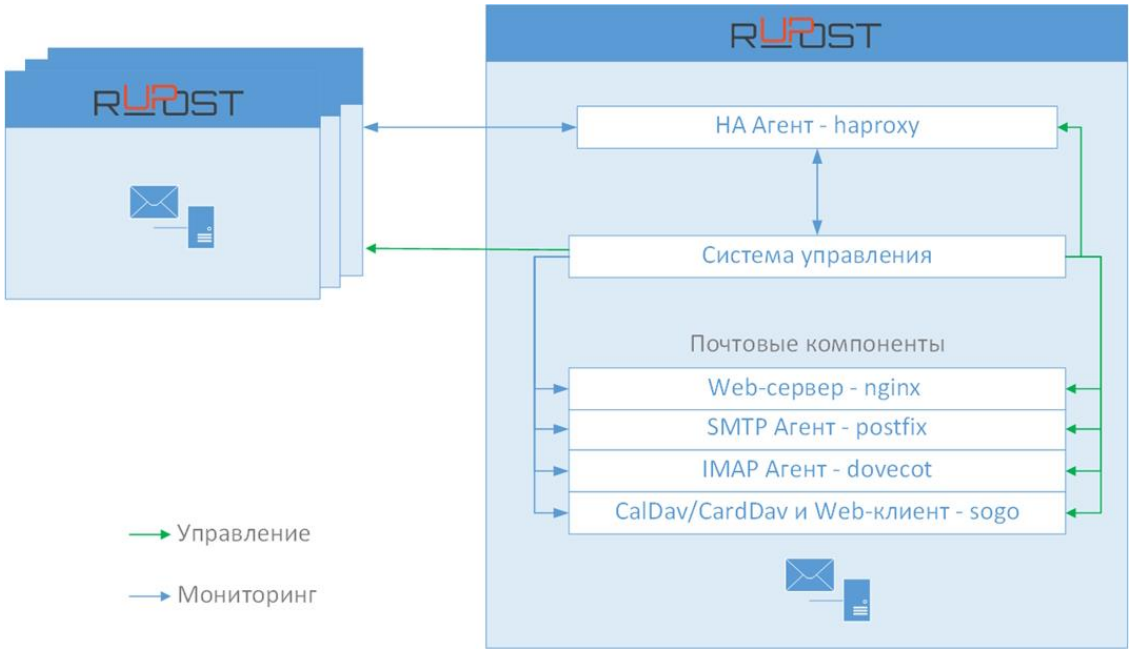
3.2. Ключевые компоненты и архитектурные особенности RuPost

При разработке архитектуры и выборе технологий RuPost, с учетом требований высокой доступности, использовался накопленный индустрией опыт применения открытых почтовых решений и практика организации высокодоступных хранилищ информации на Linux. Также, уделялось серьезное внимание особенностям развития архитектуры Exchange, как лучшего образца коробочного почтового ПО, применяемого в корпоративных масштабах.

Кластер RuPost разворачивается на нескольких узлах, на каждом из которых установлен экземпляр системы. Все экземпляры системы равнозначны и включают:

- Систему управления с входящей в нее Панелью управления
- HA Агент (агент высокой доступности, high availability), обеспечивающий коммуникации с другими HA Агентами и перенаправление полезного трафика на почтовые компоненты
- Почтовые компоненты, к которым относятся:
    - терминирующий Web-сервер – nginx
    - SMTP Агент (MTA) – postfix
    - IMAP Агент (MDA) – dovecot
    - CalDav/CardDav сервер с входящим в него почтовым Web-клиентом системы – sogo

Выбор данных компонентов в ряду альтернативных компонентов (например, существующего и популярного Exim или самостоятельно разрабатываемых IMAP-решений) обусловлен их фактическими позициями на рынке как *стандартов де-факто для организации корпоративных почтовых систем*.



Узел кластера доступен для управления и мониторинга, когда на нем функционирует, как минимум, Система управления. При этом, все почтовые компоненты (вместе с терминирующим их Web-сервером) работают как единое целое. То есть, при остановке любого из этих компонентов, останавливаются они все, причем, вне зависимости от того останавливаются ли они явно администратором или останавливается какой-либо компонент при наличии тех или иных сбоев. Почтовый трафик с узла, где почтовые компоненты остановлены, автоматически перенаправляется на другие активные узлы кластера почтовой системы. Такое решение, кроме повышения надежности, также, позволяет избежать рассинхронизации при распределенной обработке и записи писем в почтовые ящики. Данное отказоустойчивое перенаправление почтового трафика обеспечивается Системой управления RuPost.

Кроме обеспечения отказоустойчивости, архитектура RuPost, также, позволяет обеспечить постепенное/поэтапное масштабирование системы от одного узла до необходимого числа узлов кластера с автоматическим применением одной и той же активной конфигурации RuPost без необходимости индивидуального изменения конфигураций узлов. Изменение количества узлов кластера возможно во время работы почтовой системы, при этом, число узлов кластера архитектурно не ограничено.

RuPost включает следующие средства управления хранением почтовых сообщений – **Пространство хранения** (MailSpace), **Группа ящиков** (MailBox Group) и **Хранилище почты** (MailStore).

Пространство хранения (MailSpace) - совокупность нескольких хранилищ почты (MailStore), связанных правилами репликации. Необходимо наличие хотя бы одного хранилища почты. Хранилища почты делятся по выполняемым ролям - одно из них является мастером (активное, обслуживает почту в данный момент), несколько хранилищ могут быть ведомыми (slave, "горячие" реплики мастер-хранилища) и, кроме того, может быть одно резервное (backup, "холодная" реплика) хранилище. Состояние всех slave и backup хранилищ почты постоянно синхронизируется посредством периодической односторонней репликации в направлении мастер -> slave / backup. Slave хранилища почты считаются "горячими", т.е. при сбое на мастер-хранилище возможно переключение на slave.

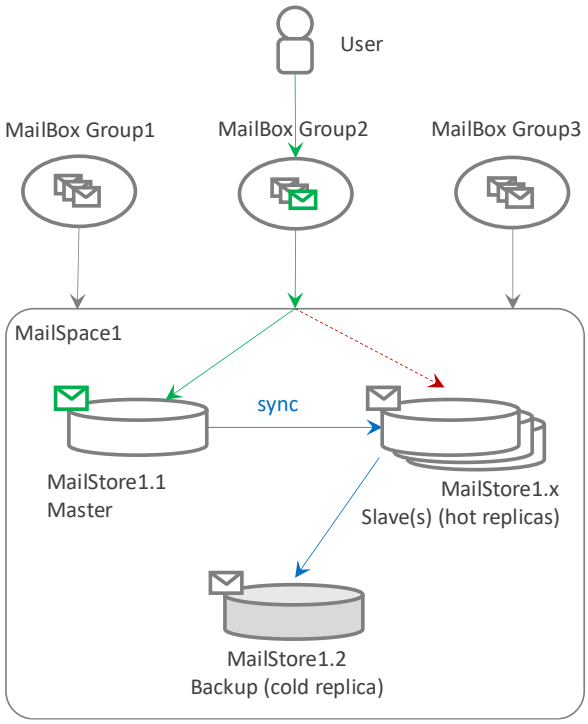
Группа ящиков (MailBox Group) — это набор почтовых ящиков, обслуживаемых одним Пространством хранения (MailSpace). Все ящики, входящие в одну Группу ящиков (MailBox Group), расположены в том Пространстве хранения (MailSpace), которое указано в свойствах этой Группы ящиков. Одно Пространство хранения может быть использовано для хранения нескольких Групп ящиков. Каждый Почтовый ящик (MailBox) принадлежит только одной Группе ящиков.

Хранилище почты или почтовое хранилище (MailStore) - набор точек монтирования NFS. Необходимо наличие хотя бы одной точки монтирования для хранения почтовых файлов в формате Maildir. В том случае, когда в Общих настройках системы установлено, что должны использоваться Архивы и/или Record Storage, то в свойствах хранилища должны быть указаны и их точки монтирования. Точки монтирования являются уникальными для всех master и slave хранилищ всех Пространств хранения. Уникальность отслеживается по полному пути - адрес NFS сервера + имя папки.

Одним из параметров Slave хранилища является “Вес”. Этот параметр задает приоритет обработки Slave хранилищ – чем выше “вес” хранилища, тем больший приоритет оно имеет. Например, для того, чтобы снизить нагрузку на master хранилище, репликация на Backup хранилище производится со Slave хранилища, имеющего минимальный “вес”. Соответственно, в рамках одного Пространства хранения, все Slave хранилища должны иметь отличающиеся значения свойства “вес”.

Резервное хранилище (Backup, "холодная" реплика). В пространство хранения может быть добавлено Backup-хранилище ("холодная" реплика), которое используется как источник данных для Системы резервного копирования (СРК). Периодичность синхронизации Backup-хранилища, в общем случае, имеет гораздо больший интервал (т.е. больше "отстает" от мастер-хранилища) чем у slave-хранилищ ("горячих" реплик). Backup-хранилище не может быть назначено мастером, т.е. на него нельзя переключить обслуживание почты.

Логические связи между сущностями, которыми оперирует RuPost, отражены на диаграмме:

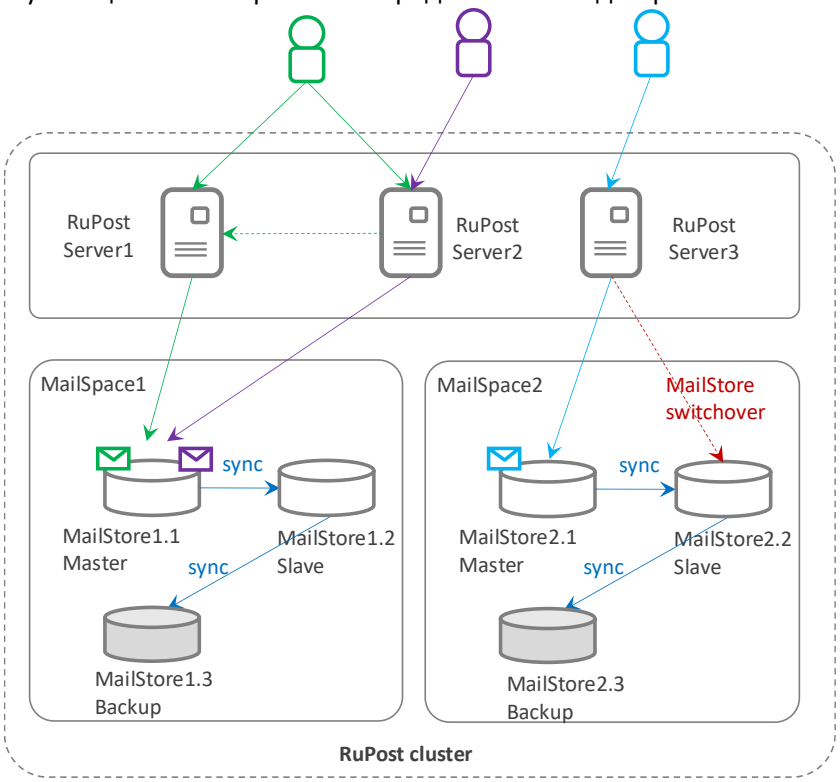


Настройка Групп почтовых ящиков, Пространств хранения и Хранилищ осуществляется как из Панели управления RuPost так и с помощью командного интерфейса (CLI).

Уникальным преимуществом представленной архитектуры RuPost является возможность одновременной работы разных пользователей с почтовыми ящиками, размещенными в одном пространстве хранения, через разные узлы почтового кластера RuPost (серверы RuPost). Такой подход существенно повышает надежность работы кластера в целом и качественно оптимизирует нагрузку на отдельные серверы обработки почты в кластере – что повышает эффективность использования инфраструктурных ресурсов и производительность почтовой системы.

MailSpace в RuPost выглядит в чем-то функционально схожим с Mailbox Database в Exchange. Однако, в отличие от Exchange, с одним и тем же MailSpace работают разные экземпляры почтового сервера RuPost. При этом MailSpace – это именно сетевое пространство хранения, монтируемое через NFS, а не специализированная СУБД как в Exchange.

Общая логика коммуникаций в кластере RuPost представлена на диаграмме:



Пользователь, попадающий в кластере RuPost через балансировку нагрузки на один из экземпляров сервера RuPost, также как и в Exchange, “привязывается” к данному серверу – узлу кластера. Однако другой пользователь, имеющий свой почтовый ящик в том же самом MailSpace, подключаясь при балансировке нагрузки к другому серверу RuPost, будет работать уже через тот сервер, на который распределен. *Это первое принципиальное отличие RuPost от Exchange, качественно снижающее риски дисбаланса нагрузки и перегруженности ресурсов отдельных узлов кластера почтовых серверов.*

Следующие подключения того же самого пользователя, попадающие через балансировку на другой сервер, проксируются им на тот сервер, к которому пользователь подключился в первый раз –

аналогично Exchange. Однако, для лучшей балансировки нагрузки, дающей более эффективное использование инфраструктурных ресурсов, в RuPost предусмотрен механизм обеспечения ребалансировки за счет периодического освобождения серверов – сброса “привязки” пользователей к конкретным серверам через штатное завершение (отключение) открытых, но не используемых – неактивных IMAP соединений. При повторном подключении пользователя по IMAP, соединение будет осуществлено на узел, имеющий наименьшее количество подключений в данный момент и, таким образом, произойдет постепенное выравнивание нагрузки по всем узлам кластера. Такое переподключение (происходит автоматически раз в сутки в ночное время – в 02:00) остается незаметным для пользователя. *Это второе принципиальное отличие RuPost от Exchange, непосредственно влияющее на сбалансированность нагрузки на почтовые серверы в кластере а, следовательно, на эффективность использования инфраструктурных ресурсов.*

Возможность работы множества серверов RuPost с одним и тем же MailSpace обеспечивается внутренней организацией MailSpace, который не является одним файлом базы данных, а представляет собой файловую структуру Maildir, где каждое сообщение является отдельным файлом.

Важно отметить, что при использовании Maildir, возможные сбои чтения-записи при работе с системами хранения влияют только на отдельные сообщения. При этом почтовый ящик пользователя остается доступным пользователю, как и все другие почтовые ящики, хранящиеся в том же самом MailSpace. Таким образом, *для обеспечения высокой доступности RuPost требуется существенно меньшее количество реплик почтовых ящиков пользователей. Это третье качественное отличие RuPost от Exchange.*

Кластер RuPost организован как группа узлов с одинаковыми экземплярами почтовых серверов, работающими в режиме Active-Active. Число узлов кластера теоретическим ограничено только ресурсами инфраструктуры и не лимитирует максимальное число узлов в отличие от Exchange, где DAG может включать в себя не более 16 узлов. *Горизонтальное масштабирование RuPost не ограничено такими рамками и обеспечивает возможность использования единого кластера для крупных инсталляций, серьезно упрощая управление крупной корпоративной почтовой системой и необходимыми ресурсами. Это четвертое важное отличие RuPost от Exchange.*

В RuPost нет необходимости создания множества MailSpaces, число которых определялось бы числом серверов в кластере, в отличие от аналогичных требований Exchange в отношении почтовых баз (в силу эксклюзивности доступа к ним). Распределение почтовых ящиков по MailSpaces происходит из организационных соображений группировки пользователей (в т.ч. с точки зрения критичности их информации) и выделения ресурсов кластерной системы хранения для разных MailSpaces.

Почтовая система является бизнес-критичной системой. Это требует обеспечения сохранности и высокой доступности самих почтовых ящиков. Как Exchange является почтовой системой, работающей только на Windows и использующей соответствующие механизмы ОС (Windows Storage Spaces, рекомендуемую файловую систему ReFS), RuPost ориентируется на механизмы управления высокой доступностью файлов ОС Linux на базе NFS, распределенных и отказоустойчивых файловых систем и программно-определяемых хранилищ SDS (Software-Defined Storage), таких как DRBD, OCFS2, zfs и др. Эти SDS доступны в Linux как компоненты “из коробки” вместе с соответствующими средствами кластеризации и синхронизации информации (например, Corosync, PaceMaker, rsync и т.п.).

Кроме работы с почтовыми базами, представленными в виде структурированного набора файлов сообщений, **RuPost использует реляционную систему управления базами данных PostgreSQL – СУБД Tantor** (или открытый PostrgeSQL, а также другие СУБД, основанные на PostgreSQL и совместимые с ним) для хранения:

- Собственных параметров и конфигураций системы RuPost
- Пользовательских календарей, контактов и адресных книг, а также параметров доступа к почтовым ящикам

Для обеспечения высокой надежности работы RuPost включает поддержку кластера баз данных PostgreSQL, построенном с использованием решения Patroni. Особенностью кластера Patroni является автоматическое переключение на другой узел кластера при обнаружении сбоя в работе основного узла. RuPost подключается к API Patroni, получает информацию о том, какой узел является основным и переключается на работу с этим узлом.

В состав кластера Patroni входят несколько узлов (СУБД), один из которых является основным (master), а остальные – репликами (replica). Данные основного узла постоянно копируются на все реплики (в синхронном или асинхронном режиме) и, таким образом, при сбое на основном узле возможно оперативное переключение на реплику (назначение одной из реплик основным узлом).

Источники

RuPost. Руководство по установке и конфигурированию.
https://www.rupost.ru/#documents

3.3. Инфраструктура почтового хранилища RuPost

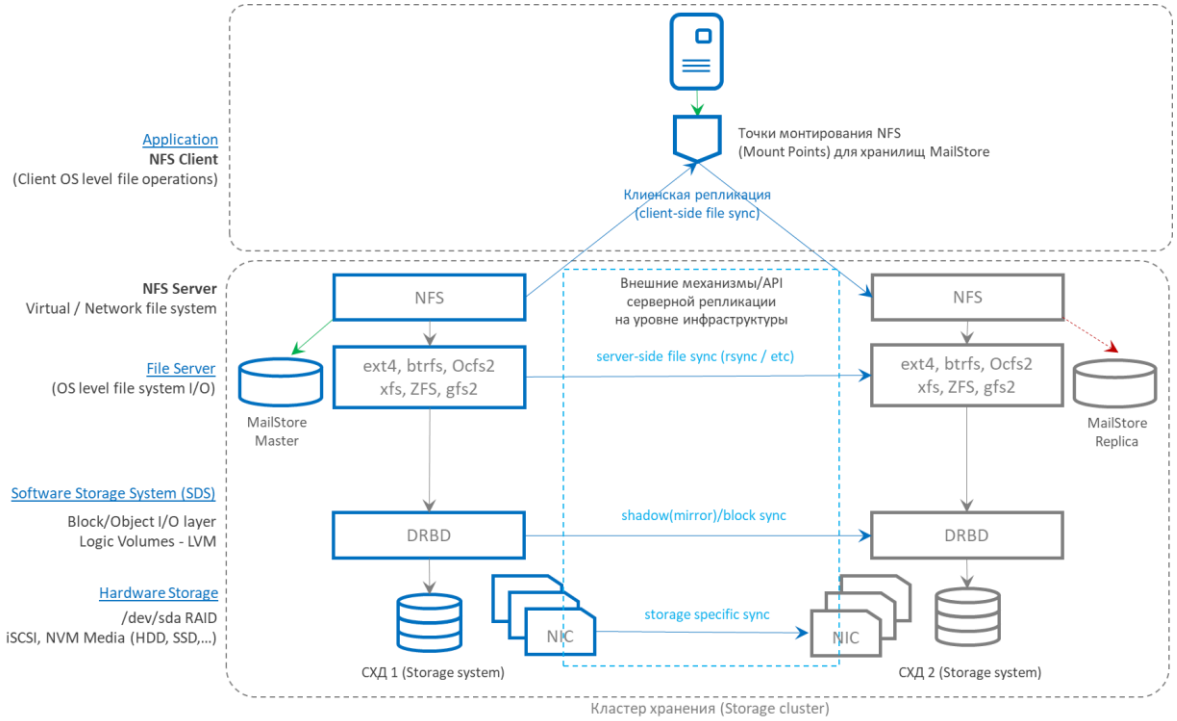
RuPost спроектирован и разработан с учетом обширной мировой практики построения кластерного отказоустойчивого хранилища Linux на базе сетевых файловых систем и протокола удаленного доступа к файлам - NFS.

Современные операционные системы Linux содержат многоуровневый стек работы с физическими устройствами хранения. На каждом из этих уровней система оперирует разными абстракциями. При этом сама ОС Linux по своей природе является ориентированной на сущность “файл”. Файлами почтовых сообщений и индексов оперируют почтовые компоненты RuPost. Доступ к этим файлам обеспечивается через стандартный интерфейс доступа к файлам в виде общего ресурса сетевой файловой системы Network File System (NFS). Фактически, NFS является интерфейсом, обеспечивающим прикладному ПО – почтовой системе RuPost – уровень абстракции хранения данных. Переключение (например, в случае сбоев или для проведения регламентных работ) на нужный ресурс легко реализуется на уровне точек монтирования NFS.

NFS – прикладной уровень работы с файлами, обеспечивает пользователям доступ к файлам, расположенным на удаленных компьютерах, и позволяет работать с этими файлами точно так же, как и с локальными. Например, с помощью команд операционной системы пользователи могут создавать, удалять, считывать, записывать и изменять атрибуты удаленных файлов и каталогов. Будучи распределенной по своей природе, NFS часто называют *виртуальной файловой системой*.

Пакет NFS содержит наборы команд и программ-демонов, предназначенных для работы с NFS, Службой информации о сети (NIS) и другими службами. Несмотря на то, что NFS и NIS обычно устанавливаются совместно, как один пакет, они независимы друг от друга, и их настройка и администрирование выполняются раздельно. Службы NFS реализованы по принципу клиент-сервер.

Модель функционирования NFS строится на использовании протокола вызова удаленной процедуры RPC (Remote Procedure Call). RPC позволяют процессам клиента запускать процессы сервера и выполнять с их помощью различные действия точно так же, как если бы процесс клиента выполнял запросы в своем собственном адресном пространстве. Так как сервер и клиент — это два отдельных процесса, они могут физически находиться как на одном узле, так и на разных узлах системы. NFS реализована в виде набора вызовов RPC, которые сервер выполняет по запросу клиентов. В свою очередь, серверные компоненты отвечают за обработку соответствующих RPC-запросов и обеспечивают непосредственный доступ к файлам в файловой системе на сервере хранения.



Следующий уровень абстракции – **файловая система (File System), непосредственно развернутая на серверах или системах хранения**. Файловая система обычно рассматривается как базовый инструмент ОС, обеспечивающий операции файлового ввода/вывода (File I/O), управляемого операционной системой (ОС).

Linux поддерживает множество файловых систем, таких как ext4, Btrfs, OCFS2, XFS и др.

- **ext4** является файловой системой по умолчанию в большинстве дистрибутивов Linux. ext4 обратно совместима с ext3 и ext2, что позволяет монтировать эти старые версии файловой системы с драйвером ext4. Однако, несмотря на все свои функции, ext4 не поддерживает прозрачное сжатие или дедупликацию данных. Шифрование File Based Encryption поддерживается начиная с ядра v4.1. Максимальное количество файлов = 2 в 32 степени.

- **Btrfs** — это файловая система Linux общего назначения нового поколения по сравнению с классическими файловыми системами ext. Btrfs не является преемником файловой системы ext4 по умолчанию, используемой в большинстве дистрибутивов Linux, но предлагает лучшую масштабируемость и надежность. Btrfs — это файловая система с копированием при записи (Copy-on-Write - CoW). Основное внимание уделяется отказоустойчивости, самовосстановлению и простоте администрирования. Дополнительным преимуществом btrfs является возможность дедупликации данных. Максимальное количество файлов = 2 в 64 степени.
- **OCFS2** (Oracle Cluster File System version 2) – открытая кластерная файловая система, поддерживающая разделяемое (совместное) использование между несколькими Linux-системами. Входит в поставку Linux и представляет из себя модуль ядра и userspace инструментарий для работы с файловой системой. Размер блока может варьироваться от 512 байт до 4 Kb, максимальный размер тома 16 Tb при минимальном размере кластера блоков в 4 Kb. Теоретический предел размера тома достигает 4 Pb при размере кластера блоков в 1 Mb. Обладает встроенными средствами журналирования, обеспечивающими целостность данных даже при аппаратных сбоях. Ориентирована на создание файловой инфраструктуры корпоративного уровня, применяется в Oracle Cloud Infrastructure в сочетании с iSCSI хранилищами.
- **XFS** первоначально разрабатывалась для поддержки чрезвычайно больших файловых систем. В отличие от ext4, XFS поддерживает функцию копирования при записи (Copy-on-Write - CoW). Максимальное количество файлов = 2 в 64 степени.
- **ZFS** – транзакционная файловая система созданная Sun Microsystems (приобретена Oracle), что означает, что ее состояние всегда непротиворечиво. Поддерживает сжатие и дедупликацию. Для использования нескольких устройств и обеспечения избыточности данных была введена концепция диспетчера томов, обеспечивающая представление нескольких устройств в виде одного устройства, чтобы исключить необходимость изменения файловых систем. Oracle поставляет специализированное корпоративное хранилище ZFS Storage Appliance, использующее эту файловую систему.
- **GFS2** (Global File System 2, Colossus) - кластерная файловая система от Google, обеспечивающая одновременный доступ к общим хранилищам в кластере. GFS2 позволяет всем узлам иметь прямой одновременный доступ к одному общему хранилищу. GFS2 использует распределенные метаданные и многочисленные журналы для обеспечения возможности кластеризации в глобальных масштабах. Как и другие файловые системы. GFS2 ориентируется на внешние средства координации ресурсов и организации кворума, необходимого для целостной синхронизации данных.

Примечание: объемы хранения данных отталкиваются от исчисления в битах, а не байтах. Соответственно, при обсуждении ограничений файловых систем по размеру принято использовать термины оценки объемов с двоичной, а не десятичной приставкой – TiB тебибайт, PiB пебибайт, EiB эксбибайт, ZiB забибайт. Единицы объема данных с десятичной приставкой необходимо пересчитывать, например: 1 TB = 0.90949470177293 TiB, 1 TiB = 1.099511627776 TB.

Выбор файловой системы определяется множеством факторов – требуемыми объемами хранения, функциональными возможностями, уровнем сложности администрирования, наличием опыта

работы с файловой системой, преимуществами совместного использования с программно-определяемыми хранилищами или аппаратными СХД, применяемыми в организации.

Некоторые из файловых систем (такие как ZFS) выделяются своими дополнительными возможностями, так как охватывают и ряд аспектов более низкого уровня абстракции – например, для работы фактически уже с системными интерфейсами логических томов дисков LVM, сетевых дисков iSCSI (сетевой протокол области хранения, который определяет, как данные передаются между хост-системами и устройствами хранения) и т.п. Обычно физическое хранение файлов и, соответственно, операции чтения-записи и передачи по сети реализуются в виде блоков – фиксированных по размеру пакетов данных (например, 4Кб). Хотя файловые системы и оперируют в дополнение к блокам такими понятиями как страница (page), это, все же, группы блоков. Именно поэтому, говоря о системном уровне хранения, часто используют термин – *уровень блочного ввода-вывода – Block I/O*.

Программные решения, которые абстрагируются от конкретного аппаратного уровня, работают с дисками по стандартным интерфейсам (например, iSCSI, NVMe и т.п.) фактически **обеспечивают реализацию программных RAID – это так называемые программно-определяемые хранилища или SDS (Software-Defined Storage)**. SDS являются менее капиталоёмким решением, чем аппаратные СХД. При этом, современные SDS системы обеспечивают высокий уровень сохранности и целостности данных, в том числе обладая встроенными механизмами репликации и дедупликации блочного уровня, что делает их жизнеспособной и популярной альтернативой применению аппаратных СХД.

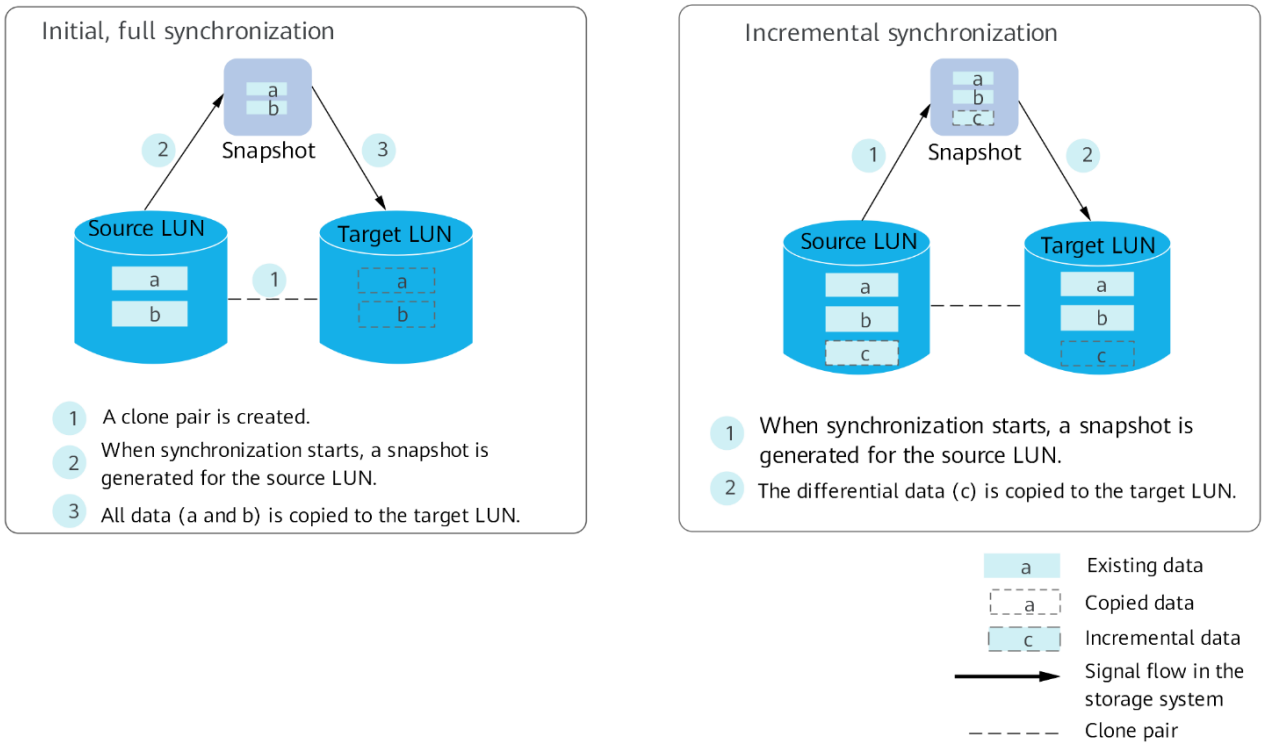
Наиболее распространенным **открытым блочным хранилищем класса SDS является DRBD**.

DRBD - Distributed Replicated Block Device – технология распределенного блочного реплицируемого устройства, являющаяся основой открытого программно-определяемого хранилища (SDS, Software-Defined Storage) LINSTOR от LINBIT. На сегодняшний день *DRBD считается открытой де факто основой для реализации высоконадежного блочного хранилища*.

В сочетании с такими кластерными файловыми системами класса OCFS2, DRBD обеспечивает высокий уровень надежности и скорости работы с файлами.

Аппаратные системы хранения данных (СХД или Storage Systems) включают специализированное программное обеспечение, эффективно работающее с блоками хранения информации на дисковых RAID-массивах, контроллерами дисков, аппаратным кешированием чтения-записи, множеством сетевых интерфейсов (Network Interface Card) и т.п. Это самый эффективный метод хранения, резервирования и синхронизации информации, в том числе в целях создания целостных реплик, размещенных в разных центрах обработки данных.

Корпоративные СХД обладают как возможностями самостоятельной настройки репликации, так и API, которые позволяют удаленным образом выполнять соответствующие команды репликации. Важно то, что такими средствами могут быть обеспечены как репликации первичные – полное клонирование заданного MailStore в контексте RuPost, так и инкрементальные – обновление и актуализация реплики MailStore (ниже в качестве примера показана диаграмма обоих видов репликации с технологией Huawei HyperClone for SAN на хранилищах OceanStor Dorado). Такая управляемая аппаратная инфраструктура хранения дает наиболее высокую производительность репликации, автоматическую отказоустойчивость хранения и защиту от сбоев и нарушения целостности данных.



Принципиальным аспектом планирования развертывания RuPost является выбор инфраструктурных возможностей для репликации данных между разными узлами инфраструктуры хранения – будь то обычные файл-серверы, полноценные SDS или СХД.

С одной стороны, RuPost включает встроенные механизмы копирования файлов между разными узлами – встроенная репликация. Однако, как показано на схеме ранее – такая репликация осуществляется на уровне приложения, то есть через клиента NFS. Это означает что все необходимые данные вначале читаются с одного узла хранения, затем передаются по сети на узел RuPost, который выступает клиентом сервера NFS, затем повторно передаются на целевой узел хранения, куда осуществляется репликация данных. Данный метод может подойти для небольшого количества малых по размерам почтовых ящиков или для передачи только произошедших изменений из небольшого активного почтового хранилища в его реплику. Альтернативой этому являются средства файловой синхронизации класса rsync, запускаемые удаленно непосредственно на серверах хранения – это обеспечивают качественно более эффективный метод копирования файлов для первичного создания реплик и их инкрементального обновления в режиме сервер-сервер, избегая избыточной передачи данных по сети как в случае с клиентом NFS, каковым является узел RuPost при использовании встроенной репликации.

Если необходимо формирование и оперативное обновление и синхронизация целостных реплик в сотни гигабайт и более, что типично для корпоративных почтовых систем, обеспечение дедупликации данных, высокой скорости копирования данных и автоматизации переключения между репликами, вплоть до репликации между различными ЦОД – требуется использование соответствующей таким требованиям инфраструктуры хранения SDS или СХД. RuPost позволяет при необходимости вызвать соответствующие команды репликации той или иной SDS и СХД удаленно с использованием ssh.

Использование SDS или СХД для почтовых хранилищ является стандартной корпоративной практикой, отвечающей базовым требованиям к хранению электронной почты в корпоративной среде.

Если сравнивать описанные инфраструктурные технологии, рекомендованные для RuPost, с **инфраструктурными технологиями, предпочтительными для развертывания Microsoft Exchange**, то можно провести определенные параллели:

- Для развертывания Exchange предпочтительно использовать технологии виртуализации дискового пространства Windows Storage Spaces, описываемые Microsoft как “технология Windows Server концептуально схожая с отказоустойчивым массивом дисков (RAID), реализованная на программном уровне”, то есть фактически программно-определяемое хранилище SDS.
- Вместе с Windows Storage Spaces предполагается использование файловой системы Resilient File System (ReFS), позиционируемой как “предназначенную для повышения доступности данных, эффективного масштабирования больших наборов данных в различных рабочих нагрузках и обеспечения целостности данных с устойчивостью к повреждениям” – частичный аналог развитых файловых систем Linux класса btrfs, zfs и других, хотя и не полностью самостоятельно реализующих их возможности (например, дедупликация, сжатие) и требующий Storage Spaces для реализации более устойчивых сценариев, как создание зеркалирование при записи (copy-on-write snapshots или shadow mirroring).

При этом, *для хранения почтовых баз данных Exchange, Microsoft рекомендует использование аппаратных СХД с непосредственным подключением их к серверам MBX как DAS (direct-attached storage) с JBOD или RAID.* Хотя использование массивов RAID и не является обязательным требованием для развертывания Exchange, “RAID является базовым (essential) компонентом при проектировании инфраструктуры хранения в Exchange (storage design) как для выделенных серверов, так и для развертываний, требующих обеспечения устойчивости к сбоям (fault tolerance)”. Более того, “Рекомендуемая конфигурация операционной системы предполагает использование технологии RAID для защиты данных от потерь. Рекомендуемые конфигурации RAID - RAID-1 или RAID-1/0, при том, что поддерживаются все типы RAID.”

В свою очередь, *RuPost максимально ориентирован на эффективное использование аппаратных СХД как сетевых хранилищ NAS или сетей хранения SAN*, с возможностью работы непосредственно с командами и API СХД для оптимизации создания и обновления реплик почтовых хранилищ. Это позволяет более эффективно использовать ресурсы корпоративных СХД. В том случае, когда СХД используется не только для RuPost, но и для других информационных систем, необходимо явное управление сетевыми настройками (зарезервированная полоса пропускания) и гарантированными параметрами ввода-вывода (i/o r/w/replication operations) с обязательным замером реальных характеристик для релевантного планирования ресурсов и параметров развертывания почтовой системы RuPost.

Для организации почтовых хранилищ в RuPost предпочтительными являются программные или аппаратные системы хранения - SDS и СХД. Выбор той или иной системы и их сочетания обуславливается имеющейся инфраструктурой, в том числе наличия аппаратных средств СХД, а также заданных критериев надежности хранения и обработки данных.

Источники

DRBD – software is a distributed replicated storage system for the Linux platform. It is implemented as a kernel driver, several userspace management applications, and some shell scripts.
https://linbit.com/linbit-software-download-page-for-linstor-and-drbd-linux-driver/

The DRBD9 User's Guide
https://linbit.com/drbd-user-guide/drbd-guide-9_0-en

Внутреннее устройство DRBD: алгоритмы работы отказоустойчивого хранилища
https://habr.com/en/companies/flant/articles/733770/

Использование DRBD и OCFS2 для резервирования данных в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=321820429

Общее хранилище для pacemaker на основе DRBD в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=331351115

Создание кластера NFS+DRBD под управлением Pacemaker в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=348174091

Работа с программными RAID в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=339496355

A Simple Guide to Oracle Cluster File System (OCFS2) using iSCSI on Oracle Cloud Infrastructure
https://blogs.oracle.com/cloud-infrastructure/post/a-simple-guide-to-oracle-cluster-file-system-ocfs2-using-iscsi-on-oracle-cloud-infrastructure

Oracle: What Is ZFS?
https://docs.oracle.com/cd/E23823_01/html/819-5461/zfsover-2.html

Rsync (remote sync), is a remote and local file synchronization tool. How To Use Rsync to Sync Local and Remote
https://www.digitalocean.com/community/tutorials/how-to-use-rsync-to-sync-local-and-remote-directories

Huawei OceanStor Dorado Technical White Paper (6.3.1 HyperClone for SAN)
https://e.huawei.com/en/material/storage/all-flash-storage/79ec5e90c01d4ee3becfacb069792b85
https://support.huawei.com/enterprise/en/doc/EDOC1100253812

Microsoft Windows Server - Storage Spaces overview
https://learn.microsoft.com/en-us/windows-server/storage/storage-spaces/overview

Microsoft Windows Server - Resilient File System (ReFS) overview
https://learn.microsoft.com/en-us/windows-server/storage/refs/refs-overview

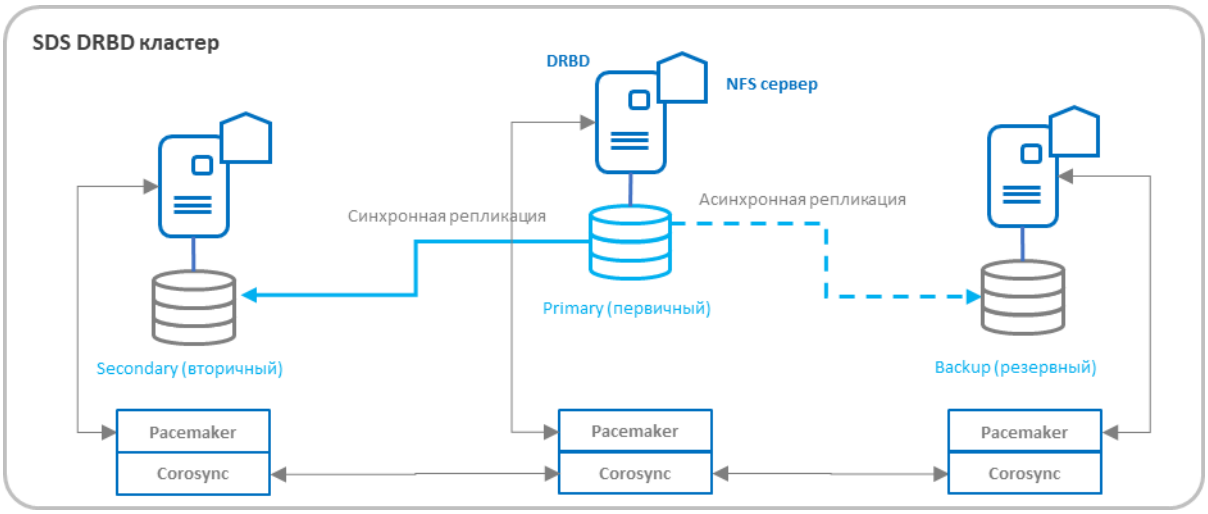
Exchange Server storage configuration options
https://learn.microsoft.com/en-us/exchange/plan-and-deploy/deployment-ref/storage-configuration

3.4. Отказоустойчивое хранилище на базе SDS DRBD

Отказоустойчивость DRBD обеспечивается дублированием блочных устройств (например, жестких дисков или томов LVM) на нескольких компьютерах (узлах), между которыми производится репликация. Репликация может выполняться в нескольких режимах или протоколах, в терминах DRBD:

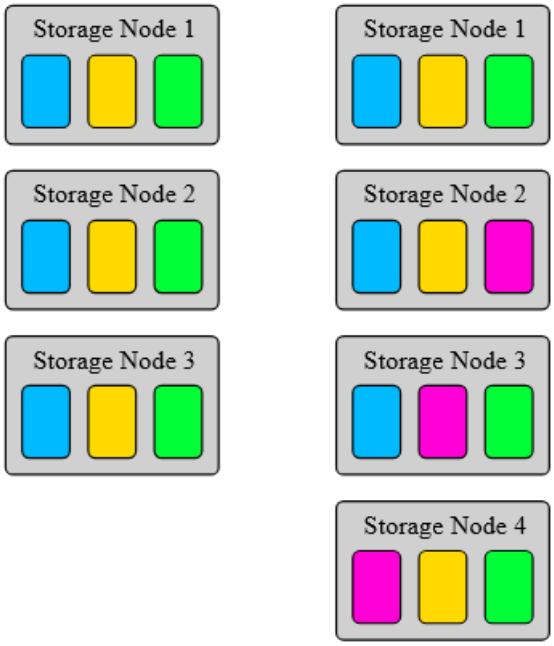
- Protocol A – асинхронная
- Protocol B – полусинхронная или репликация в памяти
- Protocol C – синхронная

В практике организации многоузлового кластерного хранилища обычно используют сочетание синхронной и асинхронной репликации. При этом, один из узлов выполняет роль первичного узла, остальные узлы выполняют роль вторичных узлов. Все обращения клиентов (чтение или изменение данных) производятся к первичному узлу. Изменения данных на первичном узле автоматически копируются (реплицируются) на вторичные узлы, на один из вторичных узлов обычно синхронно, а на последующие – асинхронно. Чтение данных всегда выполняется с первичного узла. Если первичный узел прекращает работу, то DRBD переводит вторичный узел с синхронной репликой в режим первичного. После восстановления отказавшего узла он подключается как вторичный. Последующие узлы с асинхронными репликами используются как резервные. Для корректной работы DRBD необходимо наличие ресурсов одинакового объема на всех узлах.



Для обеспечения корректного выбора первичного узла используются внешние средства обеспечения кворума – менеджер ресурсов Pacemaker и сетевой координатор Corosync.

Можно сказать, что DRBD является программной организации RAID 1+0, часто называемого также RAID10. RAID10 комбинирует зеркалирование (mirroring, RAID 1) и блочное разделение записи и чтения по разным дискам (stripping, RAID 0) для обеспечения высокой производительности. В физических системах RAID 10 представляет собой сочетание двух RAID 1, объединенных в RAID 0. Каждый диск с массива RAID 1 может быть поврежден без потери данных. RAID10 зеркало RAID 1 обеспечивает системе надежность, массив RAID 0 увеличивает производительность.



Ребалансировка данных при добавлении новых узлов кластера DRBD

При построении отказоустойчивой инфраструктуры хранения SDS на базе DRBD обычно используют кластерную файловую систему класса OCFS2, обеспечивающую дополнительный уровень надежности хранения и обработки данных.

Источники

Introduction to DRBD
https://linbit.com/drbd-user-guide/drbd-guide-9_0-en/#p-intro

Introduction to DRBD. 2.3 Replication mode
https://linbit.com/drbd-user-guide/drbd-guide-9_0-en/#s-replication-protocols

Внутреннее устройство DRBD: алгоритмы работы отказоустойчивого хранилища
https://habr.com/en/companies/flant/articles/733770/

Использование DRBD и OCFS2 для резервирования данных в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=321820429

Общее хранилище для pacemaker на основе DRBD в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=331351115

Создание кластера NFS+DRBD под управлением Pacemaker в Astra Linux
https://wiki.astralinux.ru/pages/viewpage.action?pageId=348174091

Use remote synchronous block replication on Oracle Cloud Infrastructure
https://docs.oracle.com/en/solutions/remote-synchronous-block-replication-oci/index.html

Introduction to High Availability in Ubuntu
https://ubuntu.com/server/docs/ubuntu-ha-introduction

Microsoft – Highly available NFS cluster in Ubuntu (DRBD, Corosync, Pacemaker)
https://learn.microsoft.com/en-us/samples/azure/azure-quickstart-templates/nfs-ha-cluster-ubuntu/

PaceMaker overview
https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/9/html-single/configuring_and_managing_high_availability_clusters/index#con_pacemaker-overview-overview-of-high-availability

Highly available NFS server with PaceMaker
https://www.epilis.gr/en/blog/2018/06/04/highly-available-nfs-server/

Pacemaker: Increasing System Reliability
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201406fa5.html

3.5. Отказоустойчивость баз данных

Для организации кластера PostgreSQL необходимы компоненты:

- **Оркестратор Patroni** – обеспечивает самодиагностику и автоматическую настройку, переключение между узлами кластера и потоковую репликацию данных между экземплярами базы данных
- **Координатор кластера DCS (Distributed Configuration Store)** – обеспечивает распределенное хранение конфигурационных данных кластера, ролей и статусов репликации**.**

В каждом кластере PostgreSQL на базе Patroni есть ведущий узел (лидер, primary) и ведомые узлы (реплики, standby), чьи роли могут меняться в зависимости от состояния кластера (сбой на текущем ведущем узле) и ресурсов узлов (например, перевод в ведомый узел и временный вывод, в связи с апгрейдом мощности другого узла кластера, который назначается лидером).

Patroni для определения лидера использует DCS, выбирающий узел-лидер на базе кворума. В качестве DCS могут использоваться разные системы Etcd, HashiCorp Consul, Apache ZooKeeper, и даже Kubernetes control-plane.

Большинство универсальных координаторов прикладных кластеров основаны на формировании кворума о выборе ведущего узла – лидера кластеризуемой системы на основе алгоритма *Raft*, где DCS сам по себе является многоузловым кластером с нечетным числом голосующих узлов и синхронизируемой между его узлами специализированной базой данных, содержащей информацию об узлах координируемого прикладного кластера, их доступности, статусах и текущем режиме работы и т.п. При этом, с точки зрения хранения необходимой информации, DCS обычно реализуются как хранилища типа ключ-значение (key/value).

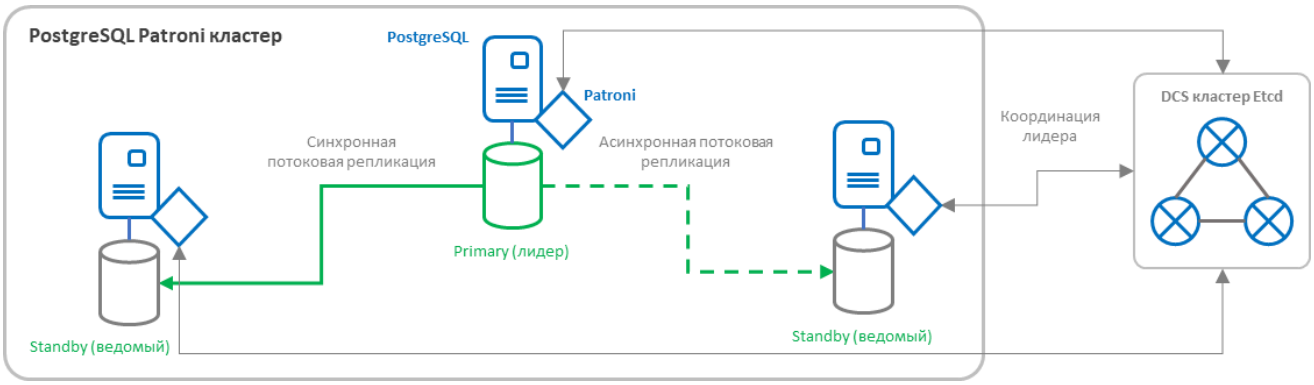
В общем случае, для организации кластера высокой доступности СУБД, необходимого для работы прикладных системам, также используются балансировщик нагрузки (HAProxy) и мультиплексоры - менеджеры пулов соединений (PgBouncer, PgPool-II или Odyssey). RuPost не требует установки дополнительных компонентов и берет на себя все функции по работе с кластеризованным сервером баз данных.

Механизмы репликации обеспечивают создание копий базы данных с ведущего узла на ведомые узлы.

В кластере PostgreSQL на базе Patroni возможны два режима потоковой репликации:

- *Синхронная* – ведущий узел (лидер) ожидает подтверждения хотя бы от одного ведомого узла перед завершением транзакции. Подтверждение успешного завершение транзакции от ведомого узла гарантирует, что данные реплицируются целостно.
- *Асинхронная* - лидер не ждет подтверждения завершения транзакции на ведомом узле, что позволяет обеспечить более высокую производительность по сравнению с синхронной репликацией. Однако, в случае сбоев нет гарантии целостности данных как при синхронной репликации.

Репликация возможна и между кластерами Patroni, включая каскадную репликацию между узлами на резервном-ведомом (standby) кластере аналогично тому, как она осуществляется в основном кластере между его узлами. Это является дополнительными средствами резервирования данных и отказоустойчивости системы за счет возможности переключения на резервный кластер.



Доступные решения для организации DCS:

- Counsul – будучи совокупностью серверных и клиентских агентов сам по себе требует развитой инфраструктуры высокой доступности управления и мониторинга HashiCorp, чтобы не оказаться единой точкой отказа (single point of failure, SPOF). Кроме того, он ограничен по рекомендациям производителя 3 или 5 голосующими узлами, а для распределенной инфраструктуры требует также зонирования с неголосующими узлами и т.п., что ограничивает его применение как для управления небольшими кластерными установками, так и наоборот – для больших распределенных систем корпоративного уровня.
- ZooKeeper обладает достаточно ограниченным API и гибкостью модели хранимой информации с одной стороны, а с другой - написан на языке java, а значит нуждается в JVM, что сказывается на его ресурсоемкости.
- Kubernetes предполагает наличие и использование отказоустойчивой инфраструктуры контейнеризации и актуален.
- Наиболее распространенным и зарекомендовавшим себя координатором с беспрецедентным уровнем устойчивости является **Etcd**. Даже Kubernetes использует Etcd в качестве Kubernetes backing store для хранения и актуализации всех данных о кластере.

Patroni обладает специальным **режимом работы Failsafe** (по умолчанию failsafe_mode выключен – выставлен в false в динамической конфигурации). Активация этого режима дает второй –

дополнительный контур обеспечения работоспособности кластера баз данных, срабатывающий в случаях:

- отказа DCS
- нарушения сетевой связанности (network partitioning) при доступе к DCS

Во втором случае – при возникновении сетевых проблем, когда нет возможности обеспечить кворум DCS, ведомый узел кластера Patroni может успешно взять на себя роль лидера, выполнив операцию leader lock update в DCS, что приведет к ситуации рассинхронизации узлов (split-brain). Режим Failsafe предназначен для предотвращения отключения лидера и позволяет при нарушении доступа к DCS (обновление блокировки лидера в DCS завершилось неудачей по другим причинам, чем несоответствие версии, значения или индекса) продолжать работать ведущему экземпляру PostgreSQL. Условием работы Filesafe является возможность прямого обращения лидера ко всем другим узлам кластера через REST-API Patroni.

Таким образом, вместе с описанной организацией кластера баз данных, **в RuPost отсутствуют недублирумые компоненты и контуры синхронизации, которые могут выступить в качестве единой точки отказа (single point of failure, SPOF), обеспечивая соответствие самым высоким требованиям к архитектуре систем высокой доступности**.

Источники

Patroni
https://patroni.readthedocs.io/en/latest/
https://app.readthedocs.org/projects/patroni/downloads/pdf/latest/

Patroni docs - DCS Failsafe Mode
https://patroni.readthedocs.io/en/latest/dcs_failsafe_mode.html

Raft consensus algorithm to manage a highly-available replicated log. What is Distributed Consensus?
https://thesecretlivesofdata.com/raft/

PostgresPro: Отказоустойчивый кластер СУБД Postgres Pro на основе Patroni
https://postgrespro.ru/clusters/patroni

Step-by-Step Guide: Configuring PostgreSQL HA with Patroni
https://bootvar.com/how-to-configure-postgresql-ha-with-patroni/

OpenText docs - High Availability PostgreSQL with Patroni
https://docs.microfocus.com/doc/128/24.3/hasqlpatroni

Streamlining Postgres Performance: Read-Only App Configuration with Patroni
https://bootvar.com/configure-read-only-apps-postgres-patroni/

Akamai Cloud - A Comparison of High Availability PostgreSQL Solutions
https://www.linode.com/docs/guides/comparison-of-high-availability-postgresql-solutions/

Cloud Architecture Center - Architectures for high availability of PostgreSQL clusters on Compute Engine
https://cloud.google.com/architecture/architectures-high-availability-postgresql-clusters-compute-engine#comparison_between_the_ha_options

NetApp instaclustr - Patroni Deployment Patterns
https://www.socallinuxexpo.org/sites/default/files/presentations/scale-21x-patroni-deployment-patterns.pdf

Percona Distribution - High Availability in PostgreSQL with Patroni
https://docs.percona.com/postgresql/14/solutions/high-availability.html

Patroni & etcd in High Availability Environments
https://www.crunchydata.com/blog/patroni-etcd-in-high-availability-environments

etcd
https://etcd.io/
https://github.com/etcd-io/etcd

etcd docs - Clustering Guide
https://etcd.io/docs/v3.4/op-guide/clustering/

etcd docs - How to Set Up a Demo etcd Cluster
https://etcd.io/docs/v3.5/tutorials/how-to-setup-cluster/

IBM Think – What is etcd
https://www.ibm.com/think/topics/etcd

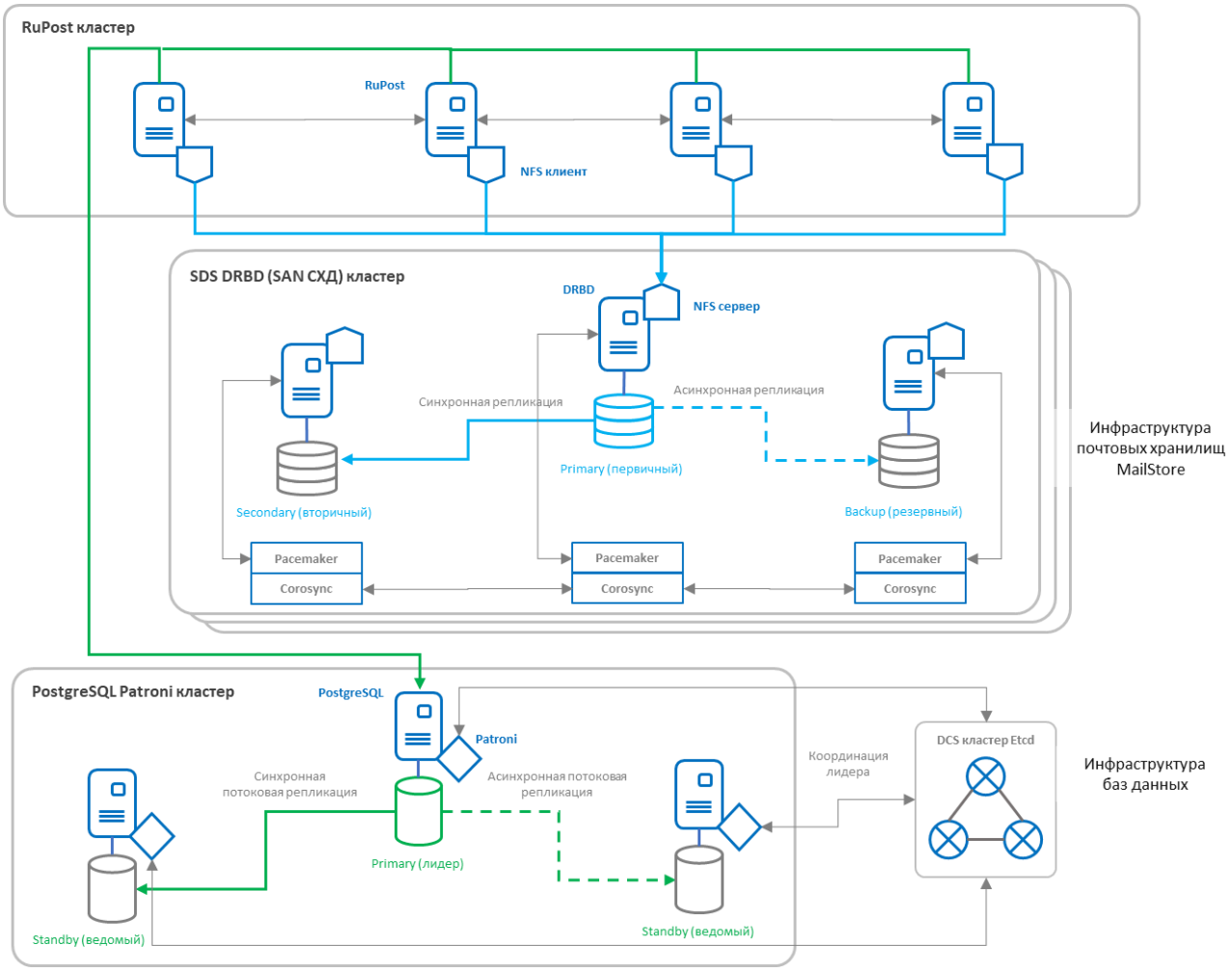
etcd versus other key-value stores
https://etcd.io/docs/v3.3/learning/why/

3.6. Обобщенная архитектура кластеризации RuPost

Обобщенная архитектура кластерной конфигурации одного сайта почтовой системы RuPost с необходимой инфраструктурой содержит три слоя:

- Кластер серверов RuPost
- Инфраструктура хранения на основе SDS или СХД (NAS, SAN)
- Кластер баз данных на основе PostgreSQL/Tantor

Отказоустойчивость каждого слоя обеспечивается присущими этому слою типовыми архитектурно-технологическими решениями.



4. Распределенные конфигурации

При рассмотрении возможных вариантов построения мультицодовых конфигураций RuPost в качестве моделей рассмотрим мультицододовые варианты кластеризации баз данных.

Типовые практики организации кластера баз данных обычно включают две модели или топологии развертывания:

1. Два ЦОД - основной и резервный. В каждом ЦОД развернут полноценный кластер БД вместе с кластером обеспечения кворума DCS. Прикладная система, работающая с кластером БД, переключается на резервный ЦОД явно – вручную, то есть в режиме switchover. Из основного в резервный ЦОД производится асинхронная репликация базы данных.

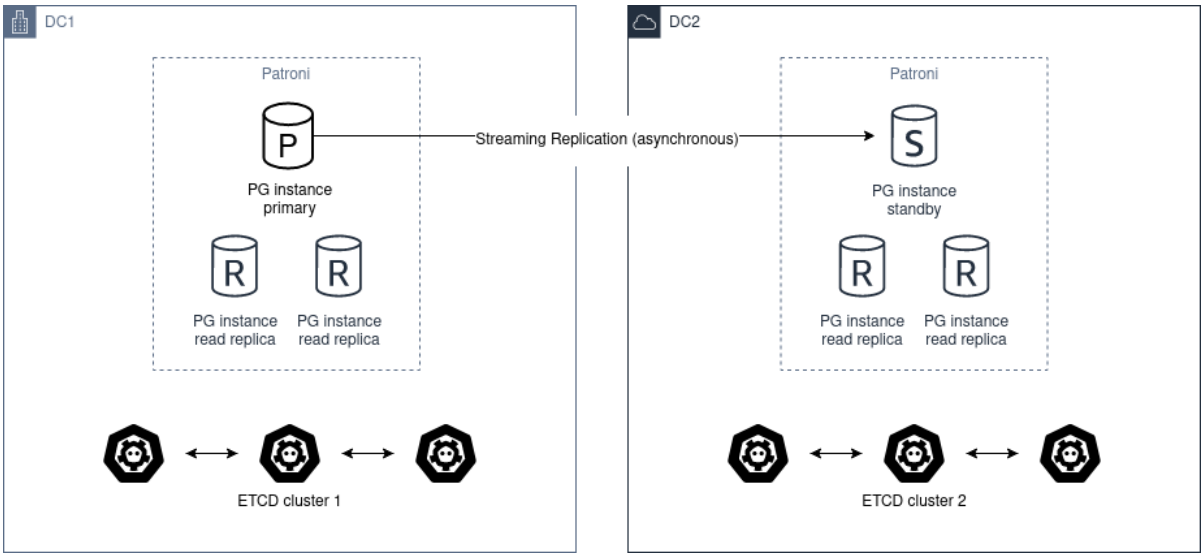


Схема активного и резервного кластеров Patroni и кластеров Etcd с асинхронной репликацией

2. Три ЦОД – “растянутый” (stretched) кластер. Во всех ЦОД размещены узлы единого кластера БД и DCS. Репликация между ЦОД осуществляется синхронно. В случае сбоя одного из узлов или в целом ЦОД, переключение производится автоматически, то есть в режиме failover.

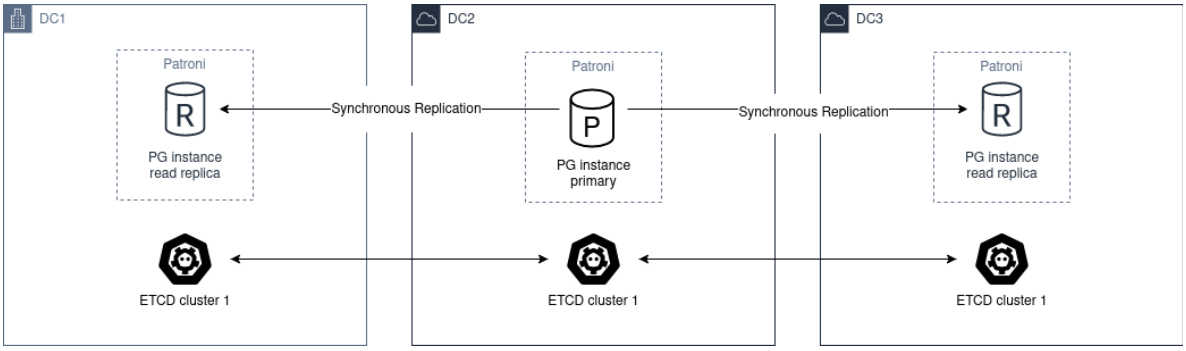


Схема кратного размещения узлов Patroni и Etcd с синхронной репликацией

При этом в первой топологии может использоваться и синхронная репликация, для минимизации рисков запаздывания данных в реплике.

В аналогичных топологиях строится и инфраструктура хранения на основе SDS.

По аналогии с рекомендованными мультицодовыми конфигурациями кластера баз данных, в случае необходимости построения **отказоустойчивой топологии RuPost**, рекомендуем применять следующие варианты:

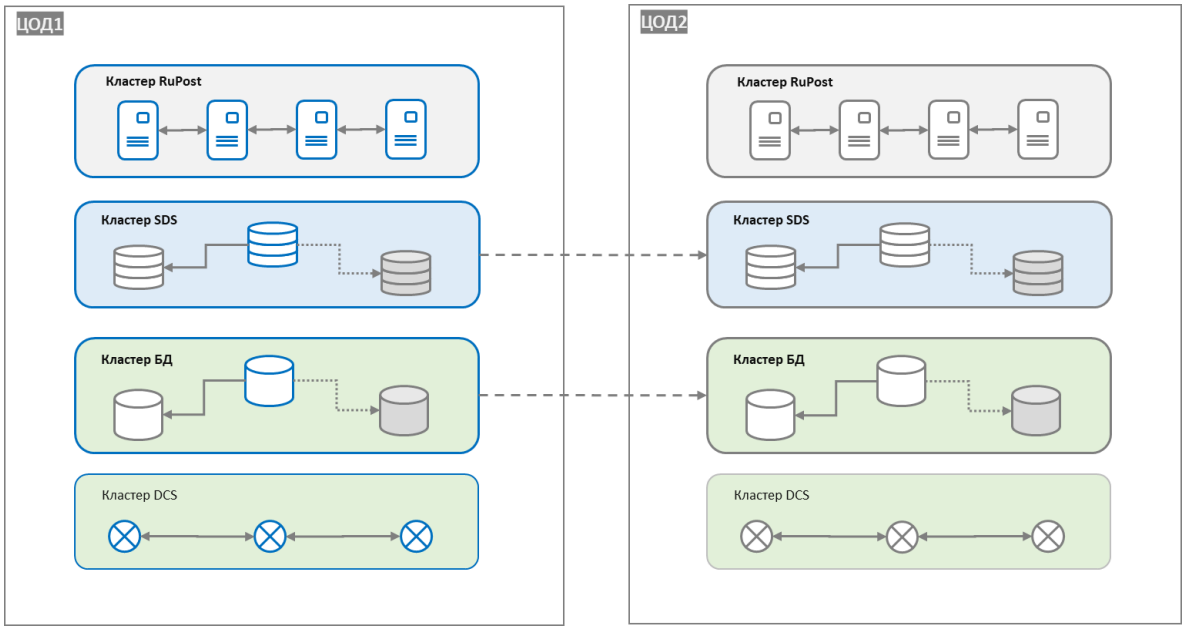
- **Резервный сайт** – дополнительный сайт RuPost, находящийся в режиме “холодного” резерва с репликацией данных основного сайта.
- **Метрокластер** – компоненты кластера распределены по нескольким площадкам в пределах одной географической локации.
- **Федеративный геокластер** или просто **геокластер** - совокупность кластеров, размещенных в разных географических локациях, при этом объединенных в единую федерацию прозрачно взаимодействующих кластеров, включая средства обеспечения сквозной маршрутизации почты между кластерами, перемещение почтовых ящиков и других данных между локациями, резервирование (репликацию) данных между локациями и т.п.

Обратите внимание - для работы RuPost нужна единая служба каталогов с репликами на каждом сайте.

4.1. Резервный сайт RuPost

Для обеспечения быстрого реагирования на потенциальные катастрофические инциденты уровня ЦОД, рекомендуется создать резервную копию сайта RuPost - продублировать необходимые кластеры в холодном (выключенном) режиме в резервном ЦОД и обеспечить постоянную репликацию данных с основного сайта. Данная **топология резервирования** является наиболее простой для мультицодовых моделей развертывания.

В каждом ЦОД развернут самостоятельный кластер RuPost с полной инфраструктурой хранения и баз данных, включая DCS. В случае отказа или недоступности всего ЦОД или отдельных инфраструктурных кластеров хранения и/или баз данных, пользователи переключаются на резервный ЦОД, на который обычно перенастраивают соответствующие записи служб DNS.

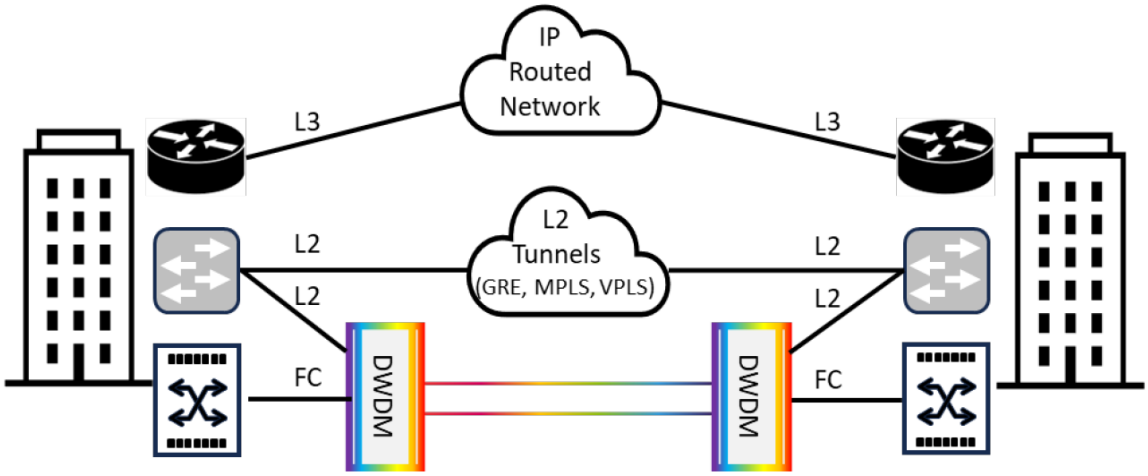


Репликация данных хранилищ и БД может осуществляться как в асинхронном, так и в синхронном режиме – в зависимости от используемой сетевой инфраструктуры обеспечения интерконнекта между ЦОД.

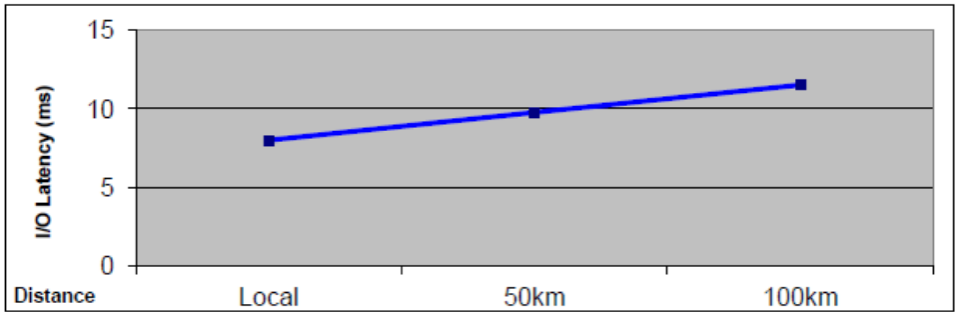
В варианте использования Геокластера, Резервный сайт является полностью самостоятельным кластером RuPost, с автоматическим конфигурированием и выполнением репликации баз данных и почтовых хранилищ с Основного сайта (см. раздел 4.3.4). При выходе из строя Основного сайта, автоматическое переключение на Резервный не происходит. На основе анализа состояния Основного сайта, администратор должен принять решение о переключении обслуживания пользователей на Резервный сайт и, при необходимости, переключить Резервный сайт в активный режим.

4.2. Метрокластер

Любая распределенная топология кластеризации, предполагающая синхронизацию данных между несколькими ЦОД, требует стабильных каналов связи между ними с соответствующей пропускной способностью и резервированием таких каналов. И здесь на первый план выходит используемая **сетевая инфраструктура**, а точнее - **качественная сетевая связность инфраструктуры обоих ЦОД**. Такая инфраструктура предполагает использование высокоскоростного интерконекта между ЦОД на базе Fibre Channel - FCIP, гарантирующего высокую скорость обмена данными и низкие сетевые задержки.



Задержки end-to-end соединений с технологиями OTN/DWDM (Optical Transport Network/Dense Wave Division Multiplexing) обычно составляют 5-25 микросекунд + скорость света. Это означает, что только в случае расстояния между ЦОД не более 50 км в принципе возможны организация отказоустойчивого распределенного сетевого хранилища Storage Area Network (SAN) на базе SDS или аппаратных СХД, сквозные транзакционные операции и синхронная репликация данных.



При планировании развертывания инфраструктур хранения и баз данных для RuPost необходимо принять во внимание оборудование и средства DPI и другие инструменты защиты периметра класса NGFW, а также маршрутизации между сегментами сети, которые вносят дополнительные существенные сетевые задержки.

Необходимо принять во внимание, что синхронный режим обычно предъявляет более строгие требования к сетевой инфраструктуре и по резервированию, и по характеристикам передачи данных.

Источники

Patroni Documentation - HA multi datacenter
https://patroni.readthedocs.io/en/latest/ha_multi_dc.html

FCIA - Fibre Channel Data Center Interconnects: 64G FC and More
(FCIA - The Fibre Channel Industry Association (FCIA))
https://fibrechannel.org/wp-content/uploads/2024/07/FCIA-DCI-finaldraft.pdf

9 Oracle RAC and Oracle RAC One Node on Extended Distance (Stretched) Cluster
https://www.oracle.com/ch-de/a/otn/docs/database/oracle-rac-on-extended-clusters.pdf

The Impact of Network Latency on Write Performance When Using DRBD
https://linbit.com/blog/the-impact-of-network-latency-on-write-performance-when-using-drbd/

4.2.1. Два ЦОД – единый кластер RuPost и разделенные инфраструктурные кластеры

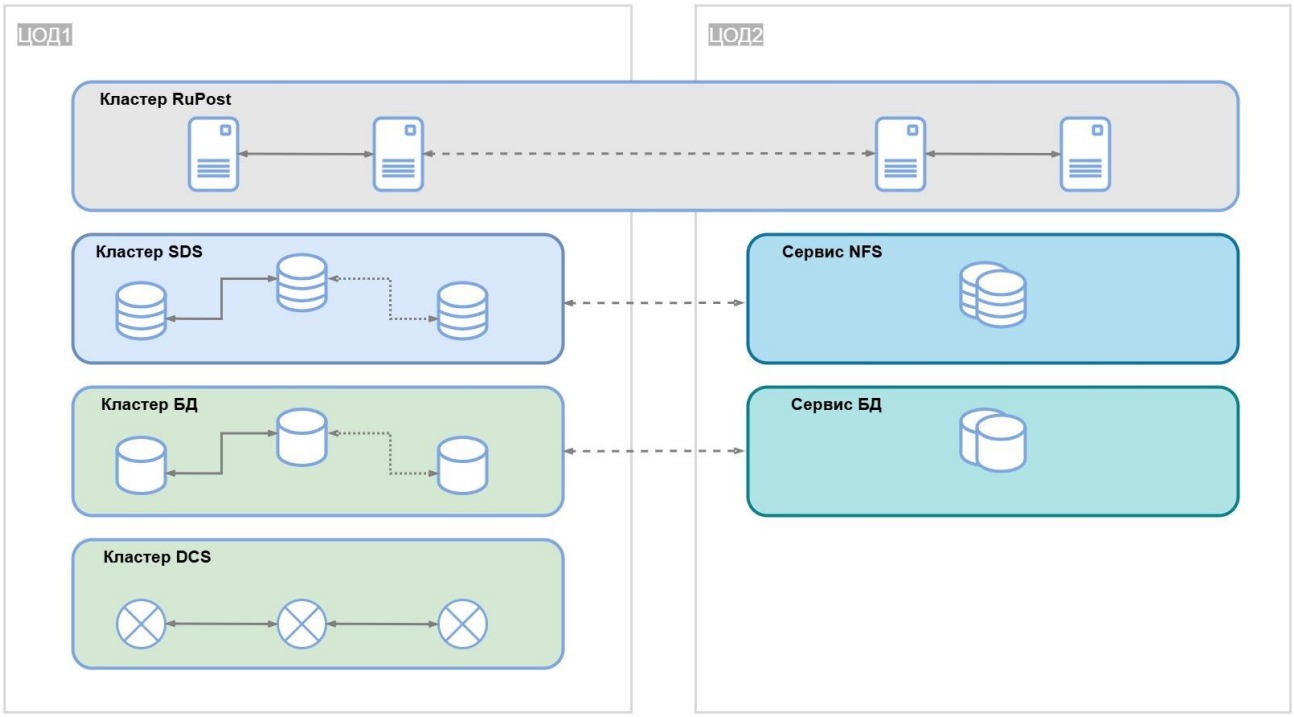
Предлагаемый к рассмотрению вариант топологии основан на схеме метрокластера и подразумевает безусловное выполнение набора требований к схеме и инфраструктуре, одними из которых являются:

- ограничение расстояния между ЦОД1 и ЦОД2 не более 50 км.;
- временная задержка в канале связи между ЦОД1 и ЦОД2, в режиме нагрузки, не более 10 мс.,
- пропускная способность канала не менее 10 Гбит.

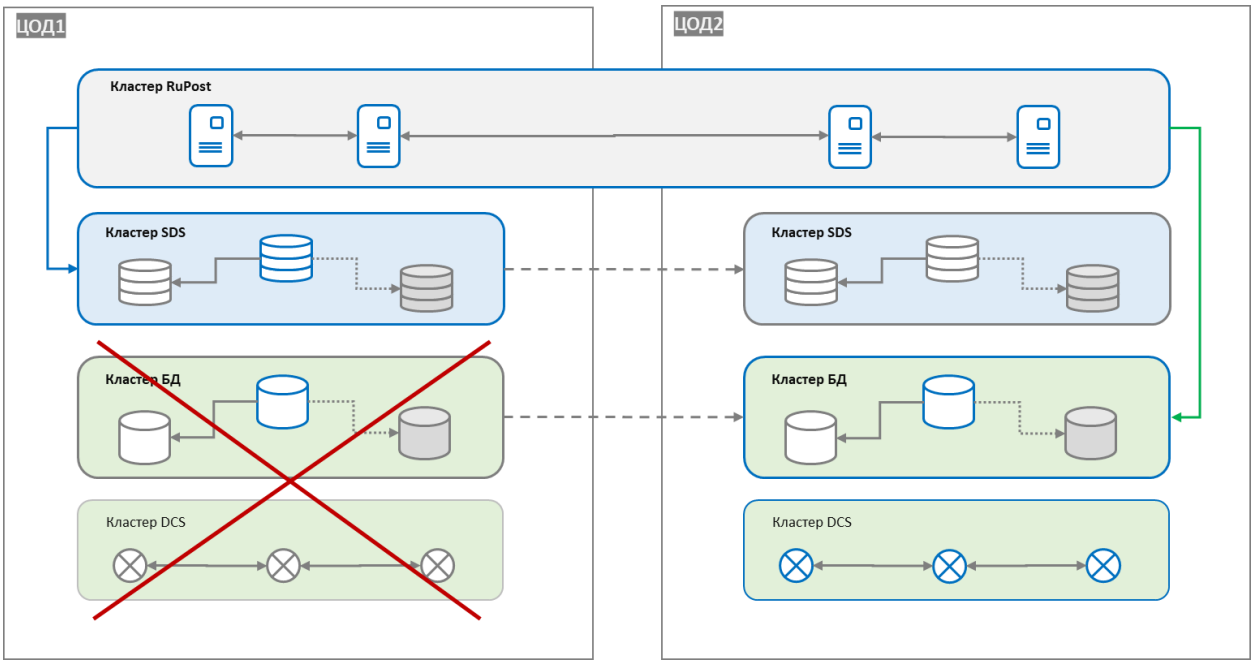
При отсутствии возможности реализации хотя бы одного из требований необходимо вернуться и использовать вариант, рассмотренный в п. 4.1 «Резервный сайт RuPost».

В случае использования асинхронной репликации в данной схеме необходимо учесть, что часть данных при переключении будет потеряна.

Преимущества данной топологии перед вариантом 4.1 выше, является более эффективное использование ресурсов единого кластера RuPost, работающего одновременно в обеих ЦОД, возможность использования в ЦОД2 не только решений на базе patroni, но и различных вариантов, СУБД как сервис (DBA as service) реализованных на базе СУБД PostgeSQL, соответствующих требованиям реализации и версии. Также возможны различные варианты реализации в ЦОД2 службы NFS как сервис (NFS as service), при соблюдении условий к версии протокола, с использованием аппаратных и программно-определяемых систем хранения данных или серверов NFS. Переключение инфраструктур хранения и БД из основного ЦОД на резервный осуществляется, также, администратором системы явно – то есть в ручном режиме switchover.



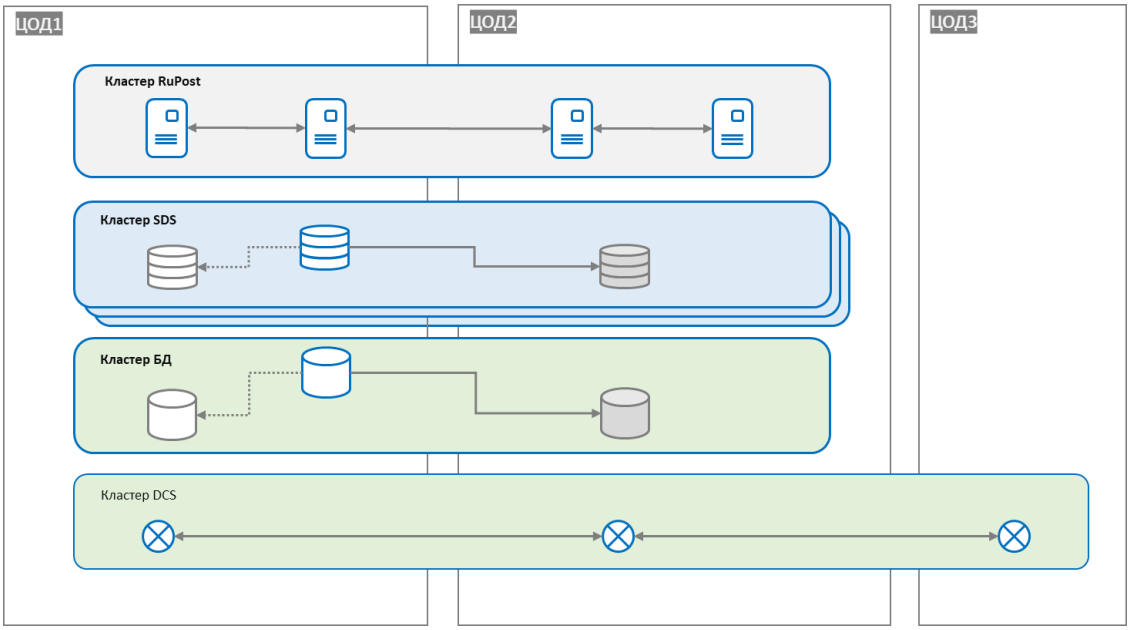
Удобством данной топологии является то, что записи DNS для узлов кластера RuPost, к которым обращаются конечные пользователи и/или другие прикладные системы, не потребуют изменений при переключении на ту или иную инфраструктуру в резервном ЦОД. В этой топологии возможным становится сценарий, когда, при выходе из строя одного из кластеров инфраструктуры хранения или баз данных в первом ЦОД, возможно подключение резервного кластера инфраструктуры из второго ЦОД. Такое переключение осуществляется администратором вручную (switchover).



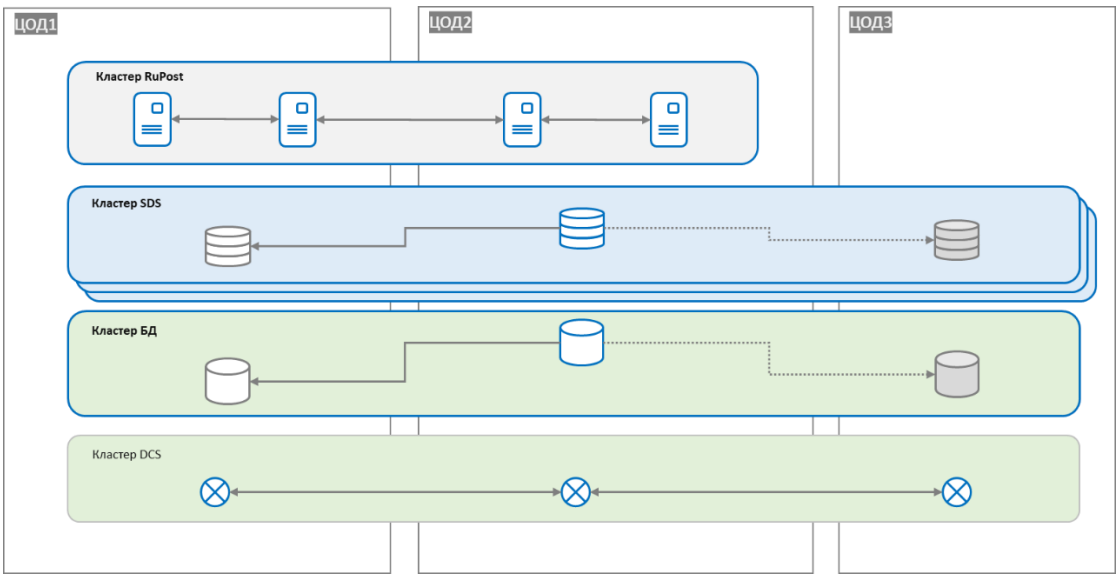
4.2.2. Полностью распределенная система

Топологии RuPost, описанные в разделах 4.1 и 4.2.1 предполагают только явный - ручной сценарий частичного или полного переключения между ЦОД – switchover. Если в перспективе рассматривается и

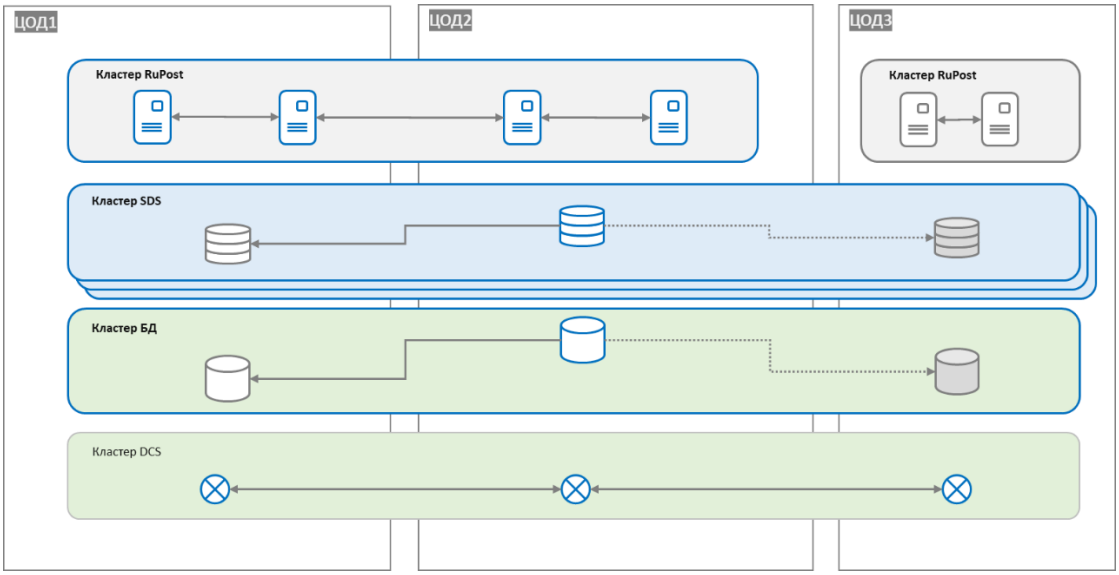
сценарий автоматического переключения между ЦОД – failover, то необходимо рассматривать системы как “растянутую” на оба ЦОД во всех своих элементах – кластер RuPost, инфраструктура хранения и кластер баз данных.



При этом, для обеспечения кворума в кластере баз данных – то есть голосования за то, какая БД будет основной – требует третьей площадки для кластера DCS. В качестве третьей площадки может выступать третий ЦОД или облако. В случае наличия достаточных мощностей в третьем ЦОД, рекомендуется рассматривать полностью распределенную систему, когда не только кластер DCS задействован в третьем ЦОД, но он используется и для резервирования почтовых хранилищ и баз данных.



Такой подход обеспечивает повышенную отказоустойчивость системы в случае выхода из строя узлов RuPost, той или иной инфраструктуры или даже целиком одного из двух ЦОД сразу с резервированием данных в третьем ЦОД.



Аналогичный сценарий развертывания Microsoft Exchange, когда узлы одного и того же DAG размещены в обоих ЦОД, называется unbound и также, как и RuPost, использует один сквозной домен аутентификации (AD) и единый почтовый домен (namespace). При этом, Exchange DAG не может содержать более 16 узлов, а RuPost такого ограничения не имеет.

Данная топология является рекомендуемой моделью организации **метрокластера**, то есть **системы, распределенной по нескольким центрам обработки данных, действуя так, как будто все узлы системы и реплики данных находятся в одном месте.**

Источники

Namespace Planning in Exchange 2016
https://techcommunity.microsoft.com/blog/exchange/namespace-planning-in-exchange-2016/604072

Load Balancing in Exchange 2016
https://techcommunity.microsoft.com/blog/exchange/load-balancing-in-exchange-2016/604048

Linbit DRBD - Multisite Data Replication Over a WAN for Disaster Recovery
https://linbit.com/blog/multisite-data-replication-over-a-wan-for-disaster-recovery/

Patroni docs - DCS Failsafe Mode
https://patroni.readthedocs.io/en/latest/dcs_failsafe_mode.html

Standby cluster
https://patroni.readthedocs.io/en/latest/standby_cluster.html

Patroni docs - HA multi datacenter
https://patroni.readthedocs.io/en/latest/ha_multi_dc.html

Exchange - Switchovers and failovers
https://learn.microsoft.com/en-us/exchange/high-availability/manage-ha/switchovers-and-failovers

Exchange - Datacenter switchovers
https://learn.microsoft.com/en-us/exchange/high-availability/manage-ha/datacenter-switchovers

Oracle RAC and Oracle RAC One Node on Extended Distance (Stretched) Clusters
https://www.oracle.com/database/real-application-clusters/#rc30category-overview
https://www.oracle.com/a/otn/docs/database/oracle-rac-on-extended-clusters.pdf

SUSE Geo Clustering Guide - Challenges for Geo clusters
https://documentation.suse.com/sle-ha/15-SP6/html/SLE-HA-all/cha-ha-geo-challenges.html
SUSE Geo Clustering Guide - Conceptual overview
https://documentation.suse.com/sle-ha/15-SP6/html/SLE-HA-all/cha-ha-geo-concept.html

Use DNS Policy for Geo-Location Based Traffic Management with Primary Servers
https://learn.microsoft.com/en-us/windows-server/networking/dns/deploy/primary-geo-location

Use DNS Policy for Application Load Balancing With Geo-Location Awareness
https://learn.microsoft.com/en-us/windows-server/networking/dns/deploy/app-lb-geo

Termidesk Connect - новый российский балансировщик нагрузки для создания отказоустойчивых ИТ-сервисов и приложений
https://termidesk.ru/news/termidesk-connect-novyy-rossiyskiy-balansirovshchik-nagruzki-dlya-sozdaniya-otkazoustoychivykh-it-se/

4.3. Геокластер - географически распределенная федерация кластеров (multi-site).

Эта предпочтительная архитектура развертывания RuPost предлагает **развертывание самостоятельных кластеров RuPost в виде “сайтов” (site) в отдельных географически разнесенных ЦОД**. Данная архитектура оптимальна для организаций с большим количеством географически-распределенных региональных или дочерних структур, каждая из которых использует свой почтовый кластер - **сайт**.

Объединение кластеров в федерацию предоставляет уникальные функциональные возможности:

- Использование единого почтового домена для всей географически распределенной организации
- Возможность использования единых корпоративных справочников (глобальная адресная книга GAL)
- Маршрутизация входящих сообщений между сайтами
- Возможность организовать единую точку обработки всех исходящих сообщений
- Прозрачный перенос почтовых ящиков пользователей между разными сайтами, размещенными в разных географических локациях
- Администраторы систем с соответствующими глобальными правами могут производить проверку статуса и высокоуровневый мониторинг доступности других сайтов и переходить с одной и той же учетной записью между консолями администратора различных сайтов (требуется единая служба каталогов или обеспечен необходимый уровень доверия между разными LDAP).

Для оптимальной организации работы пользователей в федеративной геораспределенной среде используются **специальные компоненты доступа – сервер доступа WorksPad и мобильные и настольные клиентские приложения WorksPad X и Desktop X**. Использование централизованно управляемых корпоративных клиентских приложений WorksPad X на Android и iOS и Desktop X на Linux, Windows и macOS обеспечивает возможность сквозного доступа к занятости в календарях сотрудников всех региональных структур организации, а также различным корпоративным адресным книгам, календарным ресурсам, ВКС-системам, интранет-сайтам, файловым ресурсам в соответствующем сегменте сети, ЦОД и сайте почтового сервера RuPost.

Предпочтительная архитектура геокластера как географически распределенной федерации - совокупности связанных кластеров почтовой системы - включает две топологии:

- Базовая “равный к равному”
- “Звезда” с центральным сайтом

4.3.1. Принципы работы федерации

Каждый сайт RuPost развертывается самостоятельно как самодостаточный сервер или кластер серверов RuPost. Для обеспечения управления федерацией сайтов назначается **ведущий** (leading) сайт. Объединение кластеров RuPost в федерацию осуществляется последовательным добавлением (регистрацией) сайтов к федерации. Добавление сайта в федерацию производится в Панели управления на ведущем сайте, для чего для добавляемого сайта достаточно указать его основные параметры (название, адрес HTTP, адреса SMTP и пр.), при этом будет проведена проверка

возможности подключения и сайт будет добавлен в федерацию. Все записи о внесении нового сайта в федерацию автоматически распространятся по другим сайтам федерации.

При необходимости, роль ведущего сайта может быть назначена другому сайту федерации администратором с соответствующими правами (управления федерацией). С каждого сайта федерации доступна информация о статусе других сайтов федерации и признак того, какой сайт является в данный момент ведущим. В случае недоступности ведущего сайта, администратор явно назначает другой сайт ведущим. Если сайт, ранее назначенный ведущим, снова оказывается доступен, его роль ведущего автоматически снимается и к нему, наравне с другими сайтами, применяются все глобальные настройки федерации, полученные с текущего ведущего сайта.

Почтовый ящик пользователя размещается на одном из сайтов, который является “домашним” (home) сайтом этого пользователя.

Почтовый ящик создается стандартными средствами RuPost на том сайте, где он должен быть размещен. Каждый сайт содержит список всех почтовых ящиков со всех сайтов, включенных в федерацию, с признаком принадлежности ящика тому или иному сайту. При создании почтового ящика на любом сайте, информация об этом отправляется на ведущий сайт, который, в свою очередь, информирует другие сайты о принадлежности заведенного почтового ящика конкретному сайту.

При необходимости, почтовый ящик может быть перемещен с текущего сайта на любой другой сайт. После завершения переноса данных между сайтами информация о новой принадлежности ящика сайту распространяется по другим сайтам через ведущий сайт.

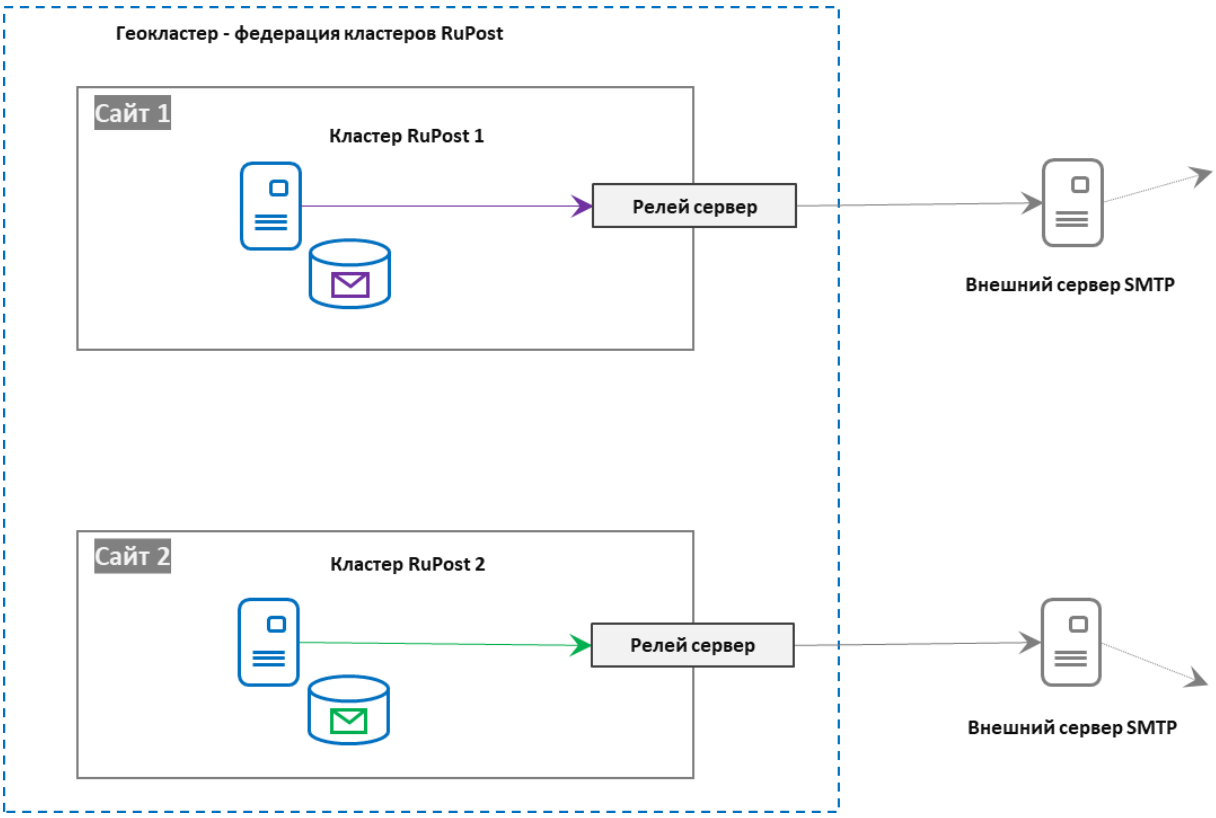
Маршрутизация входящих сообщений при попадании на сайт, который не является домашним для почтового ящика, осуществляется автоматически на основе принадлежности целевого почтового ящика тому или иному сайту.

Служба каталогов не содержит информации о принадлежности ящика сайту и может работать в режиме “только чтение” (read-only).

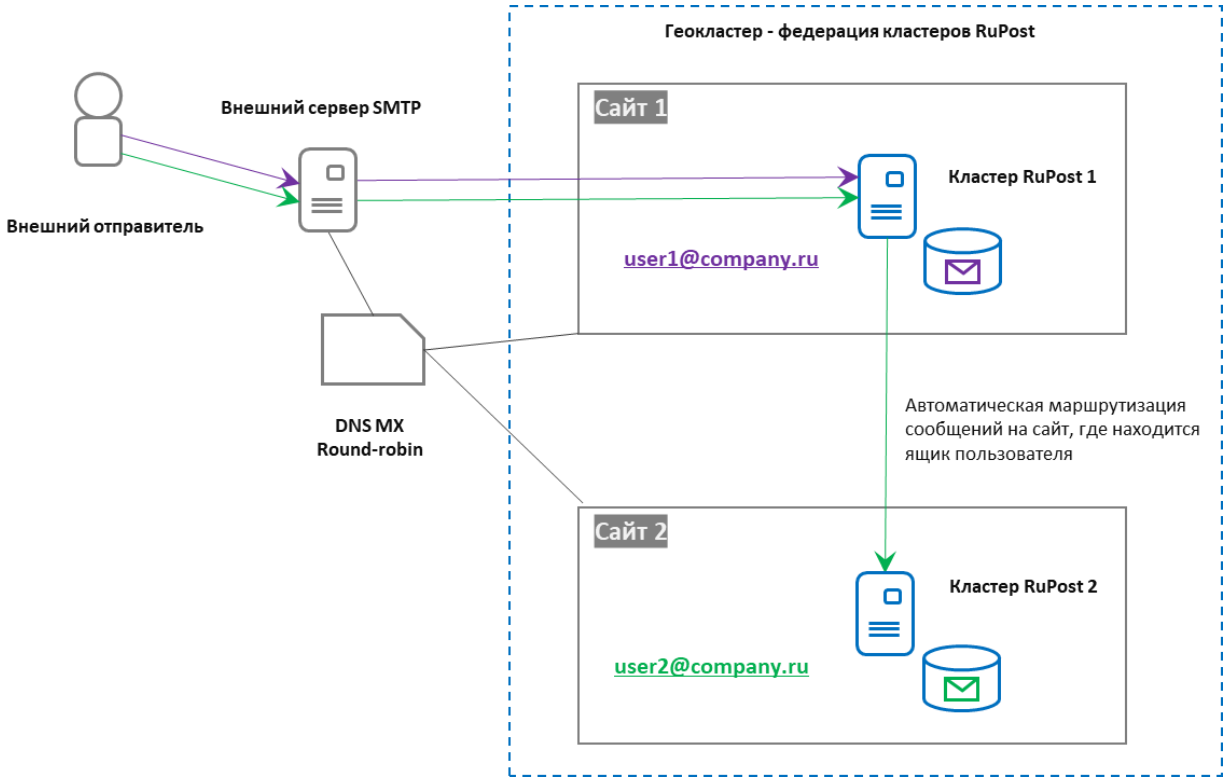
4.3.2. Базовая топология федерации

Базовая топология федерации предполагает возможность внешних коммуникаций каждого сайта вне зависимости от сетевой доступности других сайтов.

Каждый сайт для отправки сообщений использует свой релей сервер.



Входящие сообщения для любого почтового ящика могут поступать на любой сайт, не только домашний – в этом случае срабатывает автоматическая маршрутизация входящих сообщений на домашний сайт целевого почтового ящика.

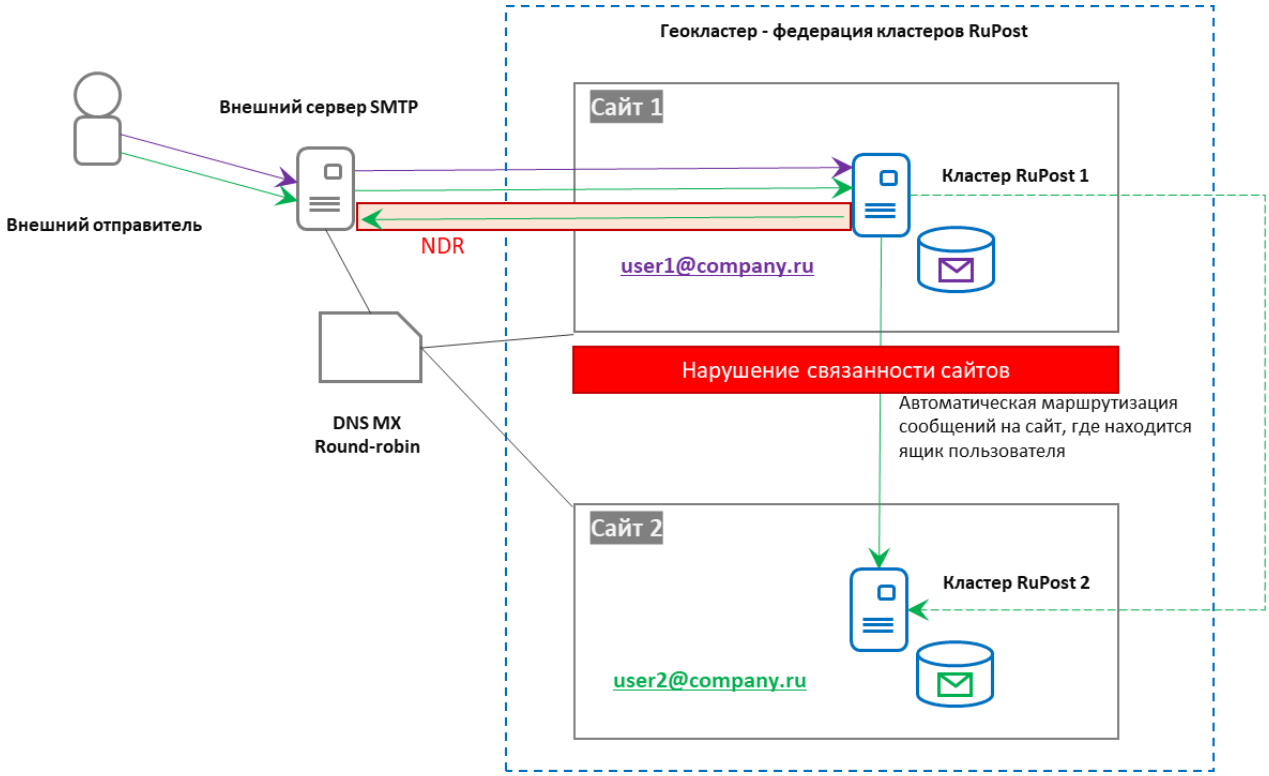


В общем случае, каждый сайт имеет, как минимум, две точки подключения:

- Внешняя (Интернет или внешний релей) – сеть, обеспечивающая доставку почты и получение почты снаружи.
- Внутренняя – сеть, обеспечивающая взаимодействие между ЦОДами.

По умолчанию, доставка почты между сайтами федерации осуществляется по внутренней сети.

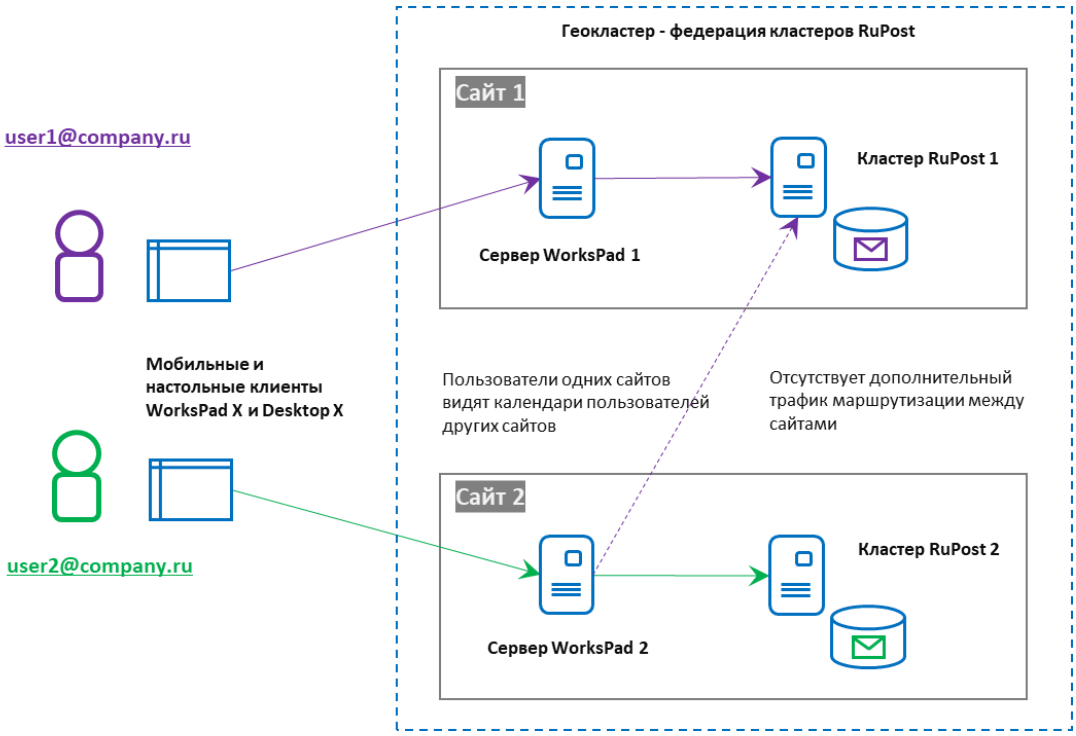
В случае нарушения связности между сайтами, каждый из сайтов продолжает работу в автономном режиме и пытается обеспечить маршрутизацию писем не напрямую, а через внешние адреса сайтов:



Отправка исходящих писем между ящиками разных сайтов, при наличии доступа “наружу” будет осуществляться через публичные каналы связи по внешним адресам сайтов без обращения к DNS MX записям, а на основании параметров, указанных в списке сайтов федерации.

При отсутствии доступа в интернет – сайт откладывает письма, адресованные пользователям с ящиками на других сайтах, до восстановления интерконнекта между сайтами или доступа в интернет. В случае длительной неработоспособности интерконнекта между сайтами сервер-отправитель smtp получит ошибку NDR (Non-Delivery Report).

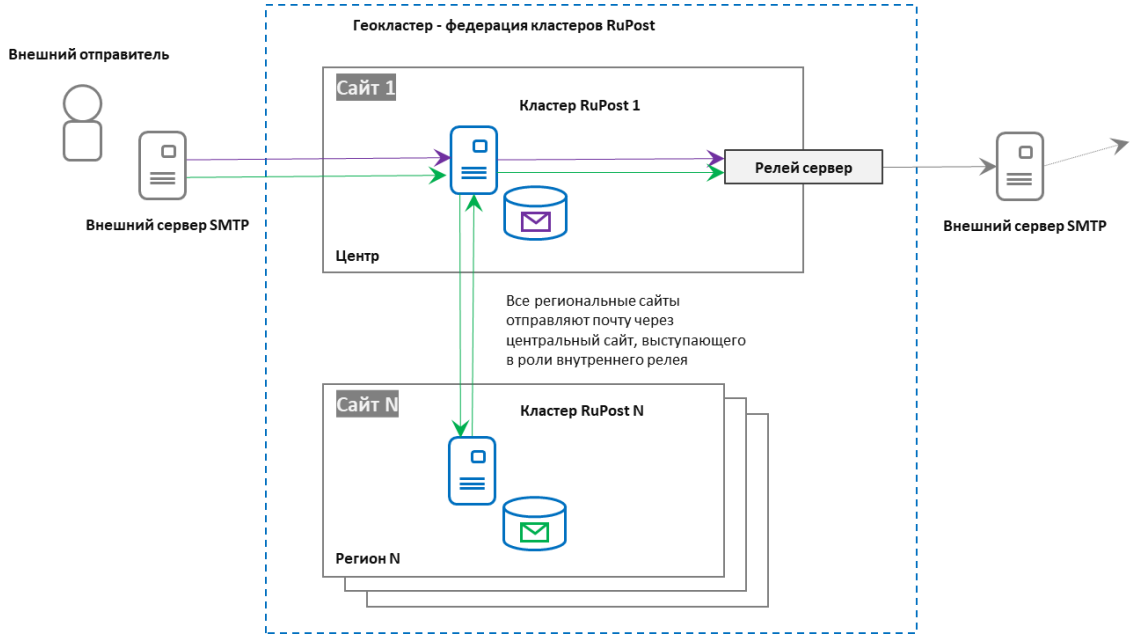
Для доступа к почте и интегрированным сервисам, корпоративные пользователи почтовой системы используют сервер доступа WorksPad и соответствующие мобильные и настольные клиенты WorksPad X и Desktop X.



Сервер WorksPad синхронизирует самих пользователей с LDAP, а карту сайтов и список принадлежности почтовых ящиков сайтам непосредственно с сервером RuPost, к которому подключен.

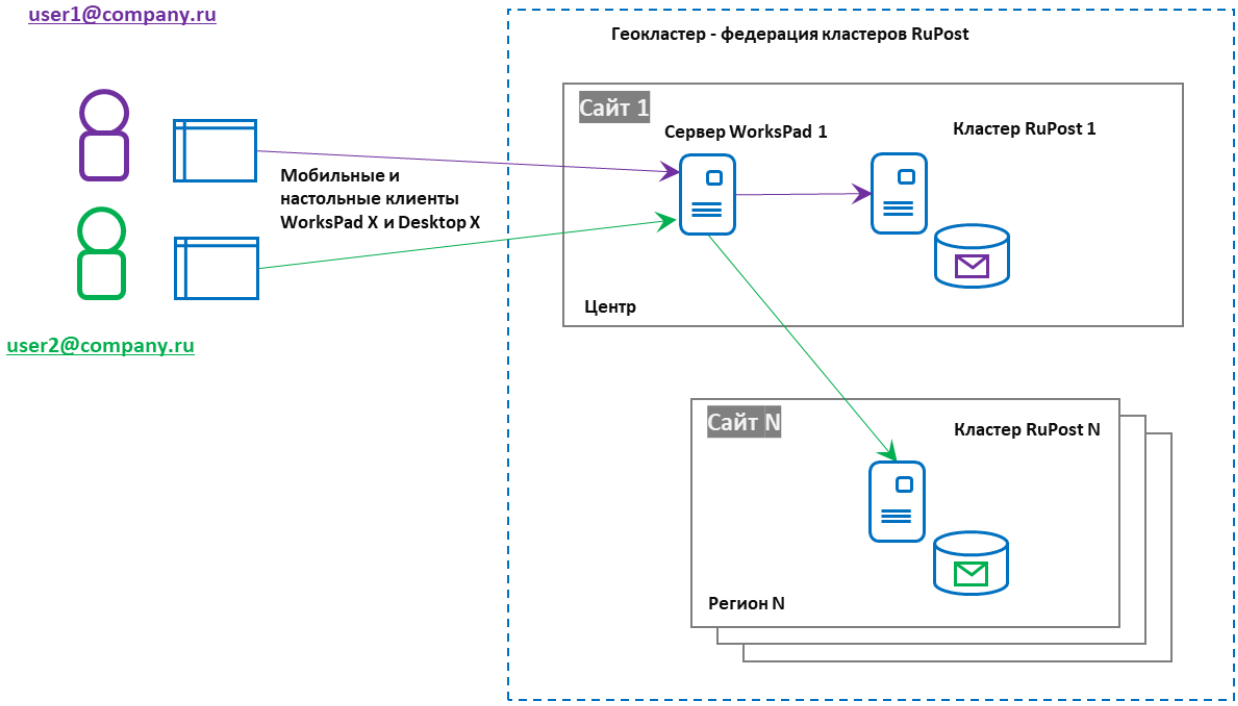
4.3.3. Топология федерации “Звезда”

В данной топологии центральный кластер RuPost - центральный сайт, выполняет функции внутреннего релей-сервера, а также прописан в DNS как сервер для всей входящей почты федерации.



Настройки внутреннего релея и правила маршрутизации почты через него автоматически распространяются на все периферийные сайты федерации.

В данной топологии доступ пользователей к почте тоже осуществляется централизованно – через центральный сайт, где размещены серверы WorksPad, которые, в свою очередь, подключаются к необходимым сайтам для обработки запросов пользователей.



Преимуществами топологии “звезда”, по сравнению с базовой топологией, с точки зрения требований информационной безопасности, является возможность применять централизованные средства контроля почтового (как исходящего, так и входящего) трафика, размещаемые на релей-серверах основного сайта.

Обеспечение отказоустойчивости каждого из сайтов решается в соответствии с выбранной для сайта архитектурой метрокластера RuPost либо с использованием резервного сайта в Геокластере (см. раздел 4.3.4 “Резервный сайт в Геокластере”). Возможные риски отказа сайта целиком могут быть снижены за счет организации резерва заданного сайта в одном из ЦОД, где расположен какой-либо из других сайтов (см. выше раздел “Резервный сайт RuPost”).

4.3.4. Резервный сайт в Геокластере

В архитектуре Microsoft Exchange, для обеспечения устойчивости сайта (**site resilience**) используется территориально-разнесенная группа серверов (DAG) – где часть серверов, входящих в группу, находятся в ЦОД1, а остальные – в ЦОД2. Критерием возможности построения такой группы является время задержки приема-передачи в сети между участниками группы – это время не должно превышать 500 миллисекунд.

Геокластер RuPost не имеет подобного ограничения – хотя, очевидно, что чем лучше качество сети между сайтами Геокластера, тем быстрее и надежнее кросс-сайтовая передача данных.

Наличие развернутого Геокластера RuPost дает дополнительные возможности для организации и управления резервными сайтами. В рамках Геокластера принципы построения резервного сайта остаются неизменными (см. выше раздел “Резервный сайт RuPost”), но, при наличии отдельного

Резервного сайта, процессы планирования, конфигурирования, развертывания и переключения на Резервный сайт существенно упрощаются и ускоряются.

Терминология

Основной сайт – сайт Геокластера RuPost, мощности которого резервируются Резервным сайтом.

Резервный сайт - сайт Геокластера RuPost, на который постоянно происходит репликация данных Основного сайта. Резервный сайт не обслуживает пользователей до тех пор, пока не будет осуществлено переключение с Основного на Резервный сайт.

Этапы работы с Резервным сайтом

1. **Планирование** – для Основного сайта, которому требуется резервирование, необходимо выбрать ЦОД, в котором будет развернут его Резервный сайт. При выборе, должно учитываться:
    a. Наличие необходимых ресурсов (вычислительная мощность, объем хранилища, ресурсы СУБД), которые могут быть использованы, в случае необходимости, для обслуживания пользователей Основного сайта.
    b. Качество сетевого подключения к Основному сайту - в первую очередь, пропускная способность канала между Основным и Резервным сайтами.

    Результат этапа – для каждого сайта в Геокластере определен резервный ЦОД.
2. **Подключение** – этот этап состоит из трех шагов:
    a. Установка Резервного сайта. Происходит полностью аналогично установке кластера RuPost только с опцией “Резервный сайт”.
    b. Назначение выбранного сайта Резервным. Администратор Геокластера в Панели управления назначает выбранный сайт Резервным для Основного сайта.
    c. Настройка Резервного сайта. Администратор резервного сайта определяет точки монтирования для всех хранилищ будущей резервной копии Основного сайта. На этом этапе полномасштабное выделение вычислительных ресурсов (узлов кластера) и ресурсов СУБД не требуется – только ресурсы, необходимые для осуществления репликации баз данных и почты.

    Результат этапа – Резервный сайт успешно сконфигурирован и начал работу по репликации данных с Основного сайта.
3. **Репликация данных** – на этом этапе производится постоянная репликация данных почтовых ящиков пользователей (MailDir) и баз данных Основного сайта.

    Результат этапа **–** данные с Основного сайта успешно реплицируются на Резервный.
4. **Переключение** – при отказе Основного сайта, администратор Резервного сайта имеет возможность включить Резервный сайт и переключить обслуживание пользователей Основного сайта на него. Так как к моменту переключения данные уже реплицированы, то переключение происходит достаточно оперативно.

Внимание!

В версии 4.2 периодичность репликации – 1 час. Соответственно, при переключении, данные за этот интервал могут быть потеряны.

Результат этапа – пользователи Основного сайта обслуживаются на Резервном сайте.

Требования к резервному сайту

Резервный сайт должен обеспечить:

При конфигурировании:

1. Учет возможностей инфраструктуры ЦОД, который планируется использовать для размещения.

В режиме репликации данных:

2. Минимальный расход вычислительных ресурсов
3. Оперативную репликацию почтовых данных пользователей.

При переключении:

4. Минимальную потерю почтовых данных пользователей.
5. Минимальное время включения.
6. Полную функциональность резервируемого сайта.

Принципы построения и функционирования резервного сайта

1. **Независимость от Основного сайта.** Резервный сайт – отдельный, полностью независимый от Основного сайта кластер RuPost с размещением в отдельном ЦОД. Соответственно, переключение обслуживания пользователей на Резервный сайта возможно даже при полном отказе ЦОД Основного сайта.

2. **Простота конфигурирования.** При подключении Резервного сайта к Основному, полная конфигурационная информация с Основного сайта копируется на Резервный. Администратор должен только осуществить минимально необходимую “привязку” сайта к локальной инфраструктуре (указать точки монтирования NFS, адреса серверов LDAP и пр.).

3. **Контроль переключения.** Переключение с Основного на Резервный сайт осуществляется администратором Резервного сайта вручную - автоматического переключения нет.

4. **Быстрое переключение.** Для переключения обслуживания пользователей Основного сайта на Резервный сайт, администратору достаточно ввести экземпляры RuPost Резервного сайта в эксплуатацию.

5. **Оперативная репликация.** Производится репликация только мастер-хранилищ почты, архивы и записи (records) - не реплицируются. Интервал репликации – 1 час.

Управление диапазоном реплицируемых данных

Так как к началу использования Резервного сайта объем почтовых данных на Основном сайте может быть довольно большим, то на Резервном сайте предусмотрена возможность восстановления информации из резервной копии почтовых данных Основного сайта перед началом репликации.

После завершения восстановления данных из резервной копии, администратор указывает начало диапазона времени для выбора почтовых сообщений, которые будут реплицированы с Основного

сайта. Сообщения, созданные раньше начала диапазона, реплицироваться не будут (т.к. они уже восстановлены из резервной копии).

Сценарий миграции сайтов между ЦОД

Резервный сайт может применяться не только для оперативного восстановления обслуживания пользователей при отказе Основного сайта, но и для миграции сайта между разными ЦОД.

Для осуществления миграции, в новом ЦОД устанавливается Резервный сайт, проходит первая репликация данных, после чего обслуживание пользователей можно переключить на Резервный сайт и удалить Основной из Геокластера.

Источники

Высокий уровень доступности и устойчивость сайта в Exchange Server
https://learn.microsoft.com/ru-ru/exchange/high-availability/high-availability

Планирование высокого уровня доступности и устойчивости сайта
https://learn.microsoft.com/ru-ru/exchange/high-availability/plan-ha

5. Средства обеспечения надежности

Система электронной почты является классическим примером системы высокой сложности с соответствующими требованиями обеспечения отказоустойчивости. В силу этого, корпоративная версия RuPost разработана в кластерной архитектуре и все три уровня кластеризации – системы управления, хранения почты и СУБД строятся в кластерном варианте.

Механизмы кластеризации системы хранения и СУБД основываются на многолетней практике обеспечения отказоустойчивости этих инфраструктурных систем.

Инновационная система управления кластером, являющаяся ядром и уникальным преимуществом RuPost, разработана и развивается с применением специальных механизмов мониторинга состояния и защиты от сбоев, которые упрощенно можно назвать средствами самодиагностики (или healthcheck, как это принято в индустрии). При этом система самодиагностики не только проверяет узлы и компоненты как системы управления кластера RuPost, но в определенном объеме и инфраструктурные системы хранения почты и базы данных.

При решении задачи обеспечения надежности для подобных систем, обязательным является наличие встроенной подсистемы мониторинга, которая является источником информации для:

- обеспечения автоматической реакции на обнаруженный сбой;
- оперативного принятия решения оператором;
- регистрации событий для последующего анализа и оптимизации работы системы.

Геокластер RuPost имеет четыре “контура” мониторинга, покрывающих все аспекты его работы. В соответствии со структурой Геокластера, обеспечивается мониторинг следующих уровней:

- Геокластер;
- Кластер;
- Инфраструктура кластера;
- Экземпляр RuPost / узел кластера.

5.1. Мониторинг Геокластера

Геокластер состоит из нескольких сайтов (кластеров RuPost), расположенных в территориально удаленных друг от друга ЦОД.

На этом уровне критичным для функционирования системы являются следующие факторы:

- **Сетевая доступность** сайтов (в соответствии с сетевой архитектурой Геокластера) – проверяется периодическими запросами API с ведущего сайта на все сайты Геокластера.

- **Консолидация информации** о состоянии Геокластера на ведущем сайте – в Панели управления Геокластером по каждому сайту отображается как информация о его сетевой доступности, так и информация о работоспособности экземпляров данного кластера.

5.2. Мониторинг кластера

На уровне отдельного сайта (кластера RuPost) критичными являются все параметры, влияющие на доставку и обработку электронной почты. Мониторинг на этом уровне осуществляется периодической проверкой доступности следующих портов почтовых компонентов для всех экземпляров кластера:

- **LMTP** / **SMTP**
- **IMAP** / **IMAPS**
- **POP3** / **POP3S**
- **HTTP** / **HTTPS**

Автоматически – при обнаружении сбоя, доступ к сбойному экземпляру полностью блокируется на уровне балансировщика и все сетевые запросы перенаправляются на работающие экземпляры. Информация о наличии сбоя, также, передается системе управления подключениями IMAP клиентов и клиенты IMAP, подключенные в данный момент к сбойному экземпляру, перераспределяются на другие экземпляры обратно пропорционально количеству уже обрабатываемых ими IMAP подключений.

5.3. Мониторинг инфраструктуры

Корректность работы системы электронной почты в большой степени зависит от качества настройки и работы инфраструктуры.

Критичными параметрами на данном уровне являются:

- **DNS** – наличие и корректность записей;
- **LDAP** – сетевая доступность серверов LDAP и необходимые права доступа для сервисной учетной записи;
- **СУБД** – сетевая доступность, возможность выполнения запросов от системного пользователя и, в случае использования кластера Patroni, статус узлов и состояние синхронизации кластера баз данных;
- **NFS** – сетевая доступность и наличие прав доступа на сервере NFS;
- **Memcached** – сетевая доступность.

Все вышеперечисленные параметры периодически контролируются, информация о результатах проверок отображается в Панели управления кластером RuPost.

5.4. Мониторинг экземпляра

Работа отдельного экземпляра (сервера) RuPost в составе кластера зависит от работы всех почтовых компонентов:

- **Балансировщик** (HAProxy)
- **Web** (Nginx)
- **MTA** (Postfix)
- **MDA** (Dovecot)
- **MUA** (SOGo)
- **Кеш запросов к базам данных** (PGPool)

Периодически проверяется статус каждого компонента и наличие критических ошибок в логах.

Автоматически - при обнаружении сбоя, делаются 3 попытки перезапуска сбойного компонента и, в случае неудачи, экземпляр выводится из эксплуатации и запускается процедура эвакуации почтовых очередей данного экземпляра на корректно работающий экземпляр. Таким образом, минимизируются ошибки обработки почты, вызванные сбоем на конкретном экземпляре RuPost в составе кластера.

Так же, периодически проверяется статус точки монтирования NFS почтовых очередей MTA.

Автоматически – при обнаружении сбоя, происходит переключение на резервную точку монтирования почтовых очередей.